John praised Mary because *he*? Implicit Causality Bias and Its Interaction with Explicit Cues in LMs

Yova Kementchedjhieva
University of Copenhagen
yova@di.ku.dk

Mark Anderson
Universidade da Coruña, CITIC
Department of CS & IT
m.anderson@udc.es

Anders Søgaard
University of Copenhagen
soegaard@di.ku.dk

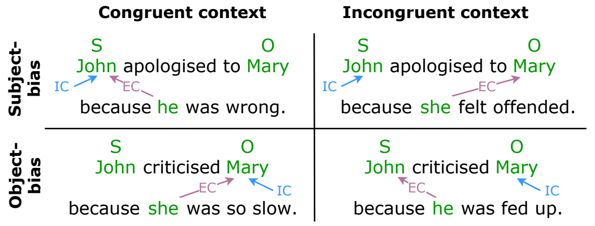
Abstract

Some interpersonal verbs can implicitly attribute causality to either their subject or their object and are therefore said to carry an implicit causality (IC) bias. Through this bias, causal links can be inferred from a narrative, aiding language comprehension. We investigate whether pre-trained language models (PLMs) encode IC bias and use it at inference time. We find that to be the case, albeit to different degrees, for three distinct PLM architectures. However, causes do not always need to be implicit—when a cause is explicitly stated in a subordinate clause, an incongruent IC bias associated with the verb in the main clause leads to a delay in human processing. We hypothesize that the temporary challenge humans face in integrating the two contradicting signals, one from the lexical semantics of the verb, one from the sentence-level semantics, would be reflected in higher error rates for models on tasks dependent on causal links. The results of our study lend support to this hypothesis, suggesting that PLMs tend to prioritize lexical patterns over higher-order signals.

1 Introduction

Recognising causal links in narrative is an integral component of language comprehension that often relies on implicit cues (Trabasso and Sperry, 1985). Pre-trained language models, which form the basis of many language processing solutions nowadays, should therefore pick up on such cues and integrate them correctly with other signals to enable accurate causal inferences in downstream tasks, including question answering and information extraction.

Psycholinguists have identified one such cue in the implicit causality bias of interpersonal verbs: some interpersonal verbs tend to implicate causality on either their subject or their object (Garvey and Caramazza, 1974). It is this bias that leads to wide agreement among subjects in psycholinguistic studies, that in a sentence like *John appreciates Mary*, the cause for appreciation likely lies with a property or action of Mary's rather than John's. Causality can also be stated explicitly, in the form of a subordinate *because* clause, for instance, and it can optionally contradict the implicit causality of the verb in the main clause. In Figure 1, we show two verbs in the context of explicit statements of causality (EC) that are either congruent or incongruent with their IC bias. Psycholinguistic studies show that congruency affects language comprehension, with human participants taking longer to identify the referent to the pronoun after *because* in incongruent contexts compared to congruent ones (Caramazza et al., 1977). The integration of the two signals costs humans extra effort, but they are eventually able to overcome the false initial expectation based on lexical semantics (i.e. IC) and form a final response that takes into consideration the full sentence-level semantics (i.e. EC). Ettinger (2020) suggests that in the context of such diverging signals, models would likely fail to integrate all signals correctly, producing a response that is consistent with the initial, shallow expectation and therefore incorrect.



John apologised to Mary because he was wrong.

John apologised to Mary because she felt offended.

John criticised Mary because she was so slow.

John criticised Mary because he was fed up.

Figure 1: Illustration of implicit causality (IC) and explicit causality (EC) in contexts where the two are congruent, and where they are incongruent.

In this work, we study a range of large transformer-based pre-trained language models (PLMs) with a focus on their awareness of IC bias

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 4859–4871
August 1–6, 2021. ©2021 Association for Computational Linguistics

and their response to stimuli of the kind shown in Figure 1. Following Ettinger (2020), we hypothesize that a language model aware of IC bias would experience interference from this signal in the context of incongruent EC, resulting in errors of judgement on a co-reference resolution task dependent on causal inference.

In a study of six PLMs from three model families: unidirectional generative, bidirectional generative and bidirectional discriminative, we find that IC bias is reliably encoded by all, but not used to an equal degree when making predictions in a controlled setting designed to test for IC awareness. In line with our hypothesis, we find that models with high IC awareness suffer an interference from IC bias in contexts of incongruent causality. We discuss these findings with reference to model type, size and amount of training data; we also draw general conclusions about the shortcomings of language models, which seem to prioritise a low-level lexical pattern (when they are aware of it in the first place) over a higher-order contextual signal.

2 Related Work

The study of the linguistic capacities of neural language models (LMs) has become especially relevant in current NLP research, where representations from PLMs feed into systems for various complex tasks, typically improving performance. Many of the testing paradigms used in psycholinguistics lend themselves well to LM analysis as they rely on a textual stimulus and a lexical response.

Linzen et al. (2016) were first to borrow from the psycholinguistic testing paradigm, in a study of the capabilities of LSTM-based models to resolve subject-verb number agreement. Goldberg (2019) adopted the psycholinguistic approach in an assessment of BERT (Devlin et al., 2019) on a number of syntactic tasks and found it to perform remarkably well on all. Hawkins et al. (2020) studied the ability of different LMs to capture human preferences as to the argument structure of English verbs.

The analysis of semantic capabilities in LMs includes studies on negative polarity in LSTM LMs (Marvin and Linzen, 2018; Jumelet and Hupkes, 2018), reasoning based on higher-order linguistic skill (Talmor et al., 2019), arithmetic and compositional semantics (Staliūnaitė and Iacobacci, 2020), stereotypic tacit assumptions and lexical priming (Misra et al., 2020; Weir et al., 2020). Many of these studies look at recent PLMs and draw mixed conclusions about the level of semantics encoded by these models. Peters et al. (2018) and Tenney et al. (2019) observed that PLMs do encode some higher-order syntactic abstractions in the higher layers (whereas lower-order syntactic information is encoded in the lower layers). However, in a comparison of contextualized and static word embeddings, Tenney et al. (2019) concluded that PLMs do not generally offer the same improvement with respect to semantics as they do for syntax.

At the crossroad of semantic analysis and psycholinguistic approaches, Ettinger (2020) introduced a suite of six psycholinguistic diagnostics for the analysis of semantic awareness in LMs. The tasks were selected based on a specific pattern observed in the response of human participants in psycholinguistic studies: an initial expectation (marked by an N400 electrophysiological response) that diverges from the final answer in a cloze task that humans offer once they have had time to fully consider the test prompt. Ettinger (2020) suggests that LMs might be "tripped up" in such contexts if they are unable to accurately integrate all the available information—she indeed found that to be the case for role-based event prediction in BERT (Devlin et al., 2019), for example. The phenomenon we study, incongruency in causality signals, has been observed to trigger a similar response in humans (Van Berkum et al., 2007) and can thus be expected to also "trip up" LMs.

Implicit causality bias was previously considered in PLM analysis by two works, both looking at how well unidirectional PLMs capture it. Upadhye et al. (2020) studied IC from the perspective of how different connectives between the main clause and the following clause (*because*, *and as a result*, full stop) affect the strength of the bias. While they did not find strong evidence for a correlation to human-based results in this respect, they did observe that in the context of connective *because* PLMs assigned lower probability to subject-referring pronouns for an object-biasing verb as compared to a subject-biasing verb. Davis and van Schijndel (2020) observed that GPT2-XL (Radford et al., 2019) encodes some level of IC bias in its representations (measured in terms of similarity between the representation of the pronoun and its two potential referents) and its decision on how to resolve a referent at prediction time is weakly influenced by that. They took the analysis one step further and looked at whether GPT2-XL uses IC information to resolve relative clause attachment, which in humans is conditioned by IC bias—no evidence was found to suggest that that was the case. Our study extends previous work on IC bias in several ways: we study both unidirectional and bidirectional models, we measure bias in the same terms as was done in psycholinguistic work and can therefore assess the correlation between the two, and we study the matter from a perspective that has not been considered before, namely the case of incongruent explicit and implicit causality.

3 Materials

Here we describe the two psycholinguistic diagnostics that we draw on: for the study of IC bias in isolation and its integration with EC. We also describe the modifications necessary to make these diagnostics suitable for PLMs.

3.1 Context-free IC Bias

Ferstl et al. (2011) studied IC bias in a context free of EC through a sentence completion task where subjects were presented with a stimulus like

(1) John praised Mary because ___

and asked to finish the sentence. Continuations were observed to start with a third person pronoun (*he* or *she*) 94.2% of the time. The researchers counted the ratio of continuations referring back to the subject of the sentence, s_{wins}, and back to the object, o_{wins}, and computed a bias score for each verb as $100 \times (s_{wins} - o_{wins})/(s_{wins} + o_{wins})$. This results in a range of -100 (verbs with extreme Object bias, hereafter O-bias) to 100 (verbs with extreme Subject bias, hereafter S-bias). The study covered 305 interpersonal English verbs with responses from 96 subjects.

In another study of IC bias, Hartshorne and Snedeker (2013) presented subjects with stimuli with a nonce ending, e.g.

(2) John praised Tim because he was a *dax*.

and asked them the question *Who do you think is a dax?* The nonce ending is meant to provide a content-free continuation that does not affect the interpretation of the ambiguous pronoun, neither semantically (as the madeup *dax* carries no meaning), nor syntactically—Hartshorne et al. (2015) conclude that explanations of the form *is a/an X* do not affect people's intuitions about who the explanation referred to.

Our approach is to use the procedure of Ferstl et al. (2011) as is to test unidirectional PLMs, as they are naturally suited to the open-end input format. Since bidirectional PLMs have been trained on complete utterances and may thus act unpredictably in an open-end context,[1] we test such models with a modification of the procedure of Hartshorne and Snedeker (2013):[2] we convert it into a cloze task with a gender mismatch between the participants, such that (3) becomes

(3) John praised Mary because ___ was a *dax*.

We adopt the mismatched-gender setting as it more closely resembles the sentence completion task in Ferstl et al. (2011). In both formats we can now identify the preferred referent by looking at the probability of pronouns *he* and *she* for the empty slot, each one referring unambiguously to only one referent. Inducing a prediction for a pronoun in the empty slot is also a more natural choice of co-reference than repeating one of the names (Holtzman et al., 2019).

In the examples shown throughout the paper, *John* and *Mary* are used as placeholders for the subject and object of the verb of interest. The choice of names to go in these slots can affect model predictions (Abdou et al., 2020), so we generate 200 variants of each stimulus, varying the names and the order between the two genders and we query the PLMs with all of them. The full procedure is described in Appendix B.

We compile a list of 200 nonce words using the 194 nonce words made available by Bangert et al. (2012), five nonce words from Cuskley et al. (2015), manually chosen to resemble English nouns, and *dax*, used in Hartshorne and Snedeker (2013). When presenting a bidirectional PLM with the aforementioned 200 variants of a stimulus, we dynamically draw a nonce word at random from this list without replacement.

The procedure described above is applied to each of the 305 verbs studied in Ferstl et al. (2011).

3.2 IC Bias in the Context of EC

Caramazza et al. (1977) tested the effect of incongruency between IC and EC using pairs of sentences built around the same verb, where one contains an explanation congruent with the verb's bias

[1] We find that a common response of generative bidirectional PLMs to stimuli like those in (1) is to predict a full-stop for the empty slot.

[2] See Appendix A.1 for an alternative we considered.

and the other contains an incongruent explanation (refer back to Figure 1 for some examples). Participants were shown one sentence at a time on a screen and asked to say out loud who the referent was to the pronoun after *because*. Caramazza et al. (1977) carried out experiments both with stimuli where the referents are of the opposite gender and where the referents are of the same gender—responses were delayed in the context of incongruent explanations as compared to congruent ones in both settings, the effect being stronger in the mismatched-gender setting.

Our approach is to adopt the mismatched-gender setting and to convert this task into a cloze task as well, an example stimulus being:

(4) John praised Mary because ___ had done well.

The stimuli used by Caramazza et al. (1977) and other related studies like Garnham et al. (1996) use only a handful of verbs (14 and 22, respectively). We therefore found it necessary to develop a more expansive dataset for the purposes of our study. Following the procedure described in Appendix C we constructed pairs of subject-referring explanations and object-referring explanations for 99 verbs, 33 strongly subject-biased verbs (bias> 65), 33 strongly object-biased verbs (bias< -65), and 33 verbs from the middle of the scale, which can be thought of as having no effect on the attribution of implicit cause. Selecting the verbs in this fashion, with large gaps between each group, allows us to see the difference between them most clearly.

Similarly to before, 200 variants of each stimulus are generated, varying the names of the referents and the order between the two genders.

4 Procedure

In this section we describe how we induce responses to the tasks described in §3 for the two experiments in this study: measuring context-free IC bias and IC bias in the context of EC. Six English PLMs are considered in this study, representative of the unidirectional generative, bidirectional generative and bidirectional discriminative paradigms in language modeling. As seen in Table 1, GPT, BERT and ELECTRA are comparable in size and training data. GPT2-M, RoBERTa-L and ELECTRA-L are included as the 'bigger siblings' to the former three models respectively, selected to resemble closely the architecture of their counterparts, while having a larger size and richer training data.[3] Comparisons can therefore be made across the three base models, on one hand, and within each pair of a base model and its larger counterpart, on the other. The three larger models are comparable in size, but not fully comparable in training data, GPT2-M being trained on only a quarter of what the other two models are trained on. In a small multilingual experiment, we also experiment with German BERT, Spanish BERT and mBERT,[4] a multilingual version of BERT.

	Work	Size	Data	Dir	Obj
English					
GPT	Radford et al. (2018)	110	16	Uni	Gen
GPT2-M	Radford et al. (2019)	345	40	Uni	Gen
BERT	Devlin et al. (2019)	110	16	Bi	Gen
RoBERTa-L	Liu et al. (2019)	355	160	Bi	Gen
ELECTRA	Clark et al. (2020)	110	16	Bi	Disc
ELECTRA-L	—"—	335	160	Bi	Disc
German BERT	-	110	12	Bi	Gen
Spanish BERT	Cañete et al. (2020)	110	20	Bi	Gen
mBERT	-	110	194	Bi	Gen

Table 1: Model properties in terms of size (number of parameters in millions), training data (size in GB), directionality (uni- or bi-directional), and token-level training objective (generative or discriminative). German BERT and mBERT are not the product of any published work, but are closely associated with Devlin et al. (2019).

We first describe the procedure for bidirectional generative PLMs which is most straightforward. Both experimental tasks can be formulated as a cloze task (see Examples (3) and (4)). We place a mask tag in the empty slot, pass the input through the model and compute the probability that the models assigns to tokens *he* and *she* for the position of the mask tag.

The procedure is equally trivial when testing unidirectional PLMs for context-free IC bias effects, where the stimuli can naturally take on an open-ended form (see Example (1)). The partial sentence is passed through the model and a probability for the relevant pronouns is computed. Measuring the effect of IC bias in the context of EC, on the other hand, cannot be performed in the next-word prediction paradigm, so for this task we instead use the unidirectional PLMs as language scorers: we

[3] For implementational details see Appendix D.

[4] The size of the training data for mBERT is not exactly known—it consists of the 100 biggest Wikipedias. English Wikipedia, as made available on HuggingFace in 2019 is 14 GB in size; with the next 99 Wikipedias being 13% the size of the English Wikipedia on average (based on number of articles), that works out to an estimate of 194GB in total.

Model	Exp	p(he) =
GPT(2)	IC	p(he $\|w_1 ... w_4$)
	IC+EC	p($w_1 ... w_4$ he $w_6 ... w_8$)
(RO)BERT(a)	both	p(MASK==he $\|w_1 ... w_4$ MASK $w_6 ... w_8$)
ELECTRA	both	p(class=O $\|w_1 ... w_4$ he $w_6 ... w_8$)

Table 2: Probing PLMs for the effect of context-free IC (IC) and IC in the context of EC (IC+EC) experimental paradigm. $w_1 ... w_4$ = *John praised Mary because*; $w_6 ... w_8$ = *was a dax / had done well* for IC and IC+EC, resp. The procedure is analogous for *she*.

create two versions of each stimulus, one with pronoun *she*, one with pronoun *he* in the empty slot, and obtain a probability for each as the average over the probabilities of all tokens in the sequence.

The discriminative model ELECTRA is trained to recognize replaced tokens in its input, i.e. for each token it computes a probability over two classes, *replaced* (R) and *original* (O). Based on the reasoning outlined in Appendix E we conclude that the more appropriate way to probe ELECTRA in our experiments is by taking the average over the probability of class O for all tokens in a sequence, instead of looking at the probability of this class for the pronoun of interest alone. In a procedure similar to the one used for GPT2-M when used as a language scorer, we present ELECTRA with two versions of each stimulus , one with *he*, one with *she* in the slot of interest, and we compute the average probability of class O for each. We formalize the handling of all model types in Table 2.

5 Exp. 1: Context-free IC Bias

With this experiment we want to determine whether English PLMs exhibit the same tendencies as humans when it comes to the IC bias of actions/states expressed with interpersonal verbs in a context free of any explicit causes. To this end, we use the materials described in §3.1 and the procedure described in §4. For any given model, the IC bias per verb is measured over the responses of the model to the set of 200 stimuli, each response processed as follows.

5.1 Measuring Bias

For a sentence with a female subject and a male object, the probability of *she* would be denoted as p^s, the probability of *he* as p^o, and $p^s > p^o$ would indicate a preference for the subject for this stimulus. Refer back to Table 2 for a summary of how these probabilities are obtained with each model.

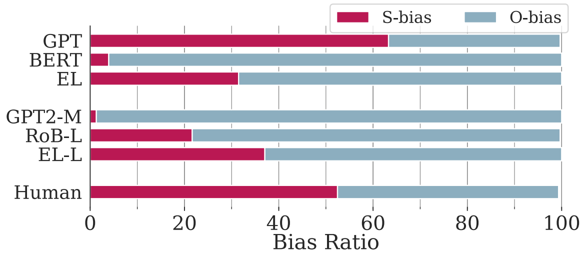


Figure 2: Ratio of S-bias verbs and O-bias verbs.

Having obtained the values p^s and p^o for each of the 200 stimuli per verb, we next calculate the bias of this verb in the following manner:

$$s_{wins} = \sum_{n \in N} (p^s_n - p^o_n) > 0 \qquad (1)$$

$$o_{wins} = \sum_{n \in N} (p^s_n - p^o_n) < 0 \qquad (2)$$

$$bias = 100 \times \left(\frac{s_{wins} - o_{wins}}{s_{wins} + o_{wins}} \right) \qquad (3)$$

where N is the set of 200 stimulus variants per verb. This metric gives us a range from -100 (extreme O-bias) to 100 (extreme S-bias), with 0 indicating an absence of any bias altogether.

5.2 Preliminary analysis

As validation of the experimental procedure, we note that the generative models, BERT, RoBERTa-L, GPT, GPT2-M, ranked one of the two vocabulary items of interest, *he* or *she*, as their top prediction at a rate of 99.0, 99.4, 99.3 and 100.0 percent, respectively. It is reassuring to see that the models behave similarly to humans in this respect, who selected *he* or *she* at a rate of 94.2% as a first token after *because* (Ferstl et al., 2011). That also indicates that the probabilities assigned to the two tokens are meaningful (Holtzman et al., 2019).

Figure 2 shows the ratio of S-bias verbs and O-bias verbs as determined by each of the models and by humans. Compared to human IC bias scores, which give an even distribution of verbs across the S-bias and O-bias classes, we see that most models show an imbalance in this respect–most notably, BERT and GPT2-M show an almost categorical preference for the object of the sentence. These trends could relate to the syntactic role of the participants (subject v. object), to their linear order, or to referent proximity. The first two factors are difficult to decouple in English, a language with a relatively fixed subject-verb-object word order. We discuss the effect of proximity in Appendix A.2.

To gain a clearer understanding of the IC bias awareness of the different models, we analyze the results of this experiment in their raw form and also with discounting for other potential sources of bias. In addition to the object bias discussed above, we include gender and choice of nonce words, which on their own did not appear to have a strong effect, but could combine with each other and with the object bias in unpredictable ways. The discounting for *p(he)*, for example, in the context of stimulus *John apologized to Mary because _ was a dax* is done by subtracting the average probability of pronoun *he* in the context of any stimulus with a male subject and word *dax* in the nonce word slot (152.5 data points on average).

5.3 Results

Table 3 quantifies the correspondence between model IC bias and human IC bias in terms of Spearman's ρ over bias scores and in terms of micro-averaged F1 score over the polarity of the IC bias (subject-bias v. object-bias). For a plot of the exact bias values see Figure 6 in Appendix F. The PLMs most affected by the discounting for other biases are BERT and GPT2-M, which also showed the strongest imbalance as observed in Figure 2. All PLMs show a significant correlation to human IC bias, although this observation has the caveat of a small dataset (only 305 data points).

Within the pairs of related models, we can say that the differences between BERT and RoBERTa-L on the one hand and ELECTRA and ELECTRA-L on the other, are small, which suggests that already with 16GB of training data and 110M parameters, these architectures reach their potential in terms of capturing and using IC bias. For the two unidirectional PLMs, we see that after discounting GPT2-M exhibits a considerably higher correlation to human IC bias scores. This may indicate that the larger size and/or richer training data of GPT2-M have enabled the model to better capture IC bias, although the correlation still remains low in absolute terms.

Comparing unidirectional PLMs to bidirectional ones, we find that the latter obtain a stronger correlation to humans scores. A similar trend holds for the F1 scores, where bidirectional models show a greater awareness of the polarity of the IC bias of verbs (especially after discounting). We refrain from making comparisons across model architectures beyond the uni- v. bidirectional dichotomy, to avoid drawing false conclusions: as we are using different procedures to induce a response from generative and discriminative models, it could be argued that a direct comparison is not methodologically robust. The discriminative models are making a binary decision over two options predefined by us, while the generative models are computing a probability distribution over hundreds of thousands of vocabulary items.

	BERT	RoB-L	GPT	GPT2-M	EL	EL-L
ρ	0.58*	0.67*	0.22*	0.22*	0.72*	0.72*
F1	0.508	0.672	0.607	0.482	0.744	0.754
ρ	0.65*	0.69*	0.23*	0.38*	0.73*	0.71*
F1	0.698	0.734	0.564	0.649	0.774	0.748
LDA	0.67*	0.58*	0.64*	0.46*	0.73*	0.67*
LR	0.71*	0.6*	0.67*	0.49*	0.75*	0.7*

Table 3: Correspondence between human- and model-induced IC bias scores (a) for model predictions, measured in terms of Spearman's ρ correlation over bias scores and F1-score over bias polarity before (rows 1 and 2) and after discounting (rows 3 and 4); and (b) for model representations (rows 5 and 6). * denotes significance at $p < 0.001$.

5.4 Further Analysis

To measure the models' sensitivity to IC bias in a perfectly comparable setting, we carry out an additional comparison on the level of representations, thus abstracting away from the top layers of the models where the differences ensue. We extract 'decontextualized' verb representation from the PLMs following the procedure described in Appendix G. Using those, we carry out two types of probes: an extrinsic one, where we train a linear regression model (LR) to map from a verb's representation to its IC bias; and an intrinsic one, where we use linear discriminant analysis (LDA) to identify the single dimension in the verb representations that is most informative of IC bias.[5] The benefit of the latter approach is that it does not add any newly trained parameters to the computation of the correlation (Torroba Hennigen et al., 2020). In both cases, the result is a vector of scalars (the values predicted by the LR, or the values of the selected dimension)—we measure the correlation between these values and human IC bias to determine how much of the latter can be recovered from the representations.

To reduce overfitting, which is inevitable with

[5] As LDA operates over a space of discrete labels, we convert the IC bias scores into 3 classes (> 0, < 0, $= 0$).

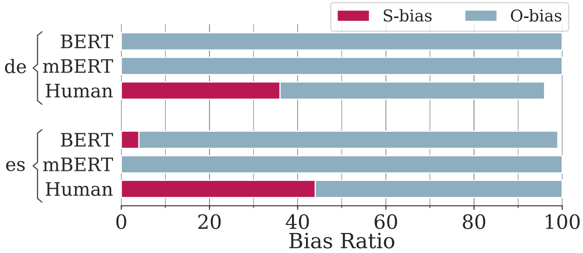


Figure 3: Ratio of S-bias and O-bias verbs in German (top) and Spanish (bottom).

305 datapoints in total and representations of 768 to 1024 dimensions, we apply PCA to the representations prior to fitting the LR and LDA models, reducing the representations to 5% of their original size. Each model (LR and LDA) is fit on a random 50% split of the data and applied on the other 50% to predict (LR) or transform (LDA). This procedure is repeated 100 times for robustness. The mean correlations are reported in the last two rows of Table 3. We see that larger models yield lower correlations than their smaller counterparts, suggesting that the former might encode IC bias in a more distributed manner than the latter.

Comparing models of equal size, we see that a similar pattern holds here as observed over the models' predictions, with the unidirectional models showing a lower correlation than the bidirectional ones, although the gap is substantially smaller in this space. It appears that unidirectional models might encode more IC bias than they exhibit at inference time. ELECTRA shows the highest correlation among the base-size models and ELECTRA-L the highest among the large models–as this comparison abstracts away from the specific objective each models uses (generative v. discriminative), we can conclude that the two ELECTRA models capture IC bias to the greatest extent out of the six PLMs studied here.

5.5 IC Bias in Other Languages

IC bias is not an English-specific phenomenon—Goikoetxea et al. (2008) obtained human judgements for 100 Spanish verbs, and Van den Hoven and Ferstl (2018) did so for 100 German verbs. Here, we probe Spanish (es) BERT, German (de) BERT and mBERT for their IC bias awareness. Details on the choice of proper nouns and nonce words are discussed in Appendix H. As seen in Figure 3 a recency/object bias is observed for the PLMs investigated here as well, so we present the results with and without discounting.

	German		Spanish	
	BERT	mBERT	BERT	mBERT
ρ	0.54*	0.23	0.13	-0.00
F1	0.600	0.600	0.540	0.560
ρ	0.51*	-0.13	0.16	-0.15
F1	0.680	0.380	0.610	0.360
LDA	0.26*	-0.0	0.09	0.08
LR	0.47*	0.02	0.12	0.03

Table 4: Correspondence between human- and model-induced IC bias in German and Spanish. For more details see the caption of Table 3.

Table 4 summarizes the results before (rows 1 and 2) and after discounting (rows 3 and 4). The poor performance of the multilingual mBERT is not surprising—Rönnqvist et al. (2019) found mBERT to be inferior to monolingual models at making a prediction for randomly masked subtokens (specifically looking at German, among other languages); and Vulić et al. (2020) found mBERT and XLM-R to both be inferior to their monolingual counterparts on probing tasks pertaining to lexical semantics.

German BERT shows a medium-strength correlation to human scores, whereas Spanish BERT shows no such correlation at all, both on the level of predictions and model representations. This observation could be attributed to the pro-drop nature of Spanish, wherein pronouns are often dropped when in subject position. This likely makes the learning of IC bias in Spanish harder for a PLM, as less evidence is available in the context to connect the explanation to its referent.

From this section, we conclude that English bidirectional PLMs reliably capture and use IC bias in their predictions. Unidirectional models encode IC bias but do not greatly rely on it at prediction time. Having established that IC bias affects the behavior of at least some PLMs, we now evaluate how these models integrate this implicit signal with more explicit signals from the sentence-level semantics.

6 Exp. 2: IC Bias in the Context of EC

With this experiment, we test the hypothesis that when the IC and EC signals converge in congruent contexts, i.e. they point to the same referent, the models would have more ease predicting the correct referent, whereas when the two signals diverge in incongruent contexts, the models would be more prone to errors. We test this hypothesis using the

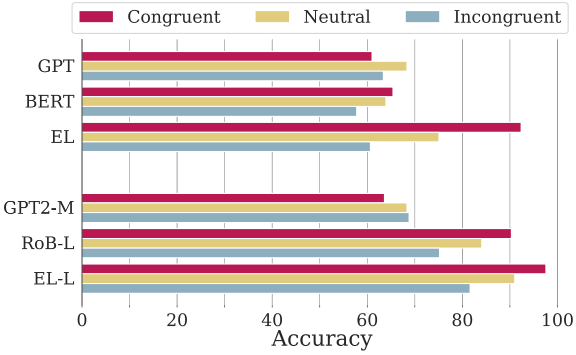


Figure 4: Accuracy on co-reference resolution over stimuli with congruent and incongruent IC and EC.

materials described in §3.2 and the procedure from §4. We present each stimulus to the models in 200 versions varied for subject and object referents. In this experiment, we do not perform discounting: unlike Experiment 1 where we wanted to gain as clear a view as possible of the level of IC bias that models exhibit, isolated from other sources of bias, here we want to see how IC bias interacts with EC, subject to any other potential sources of bias.

6.1 Results

Figure 4 shows the results from this experiment. All models are substantially better at resolving the antecedent correctly on average compared to a random baseline of 50%. Looking at the neutral stimuli as indication for the models' general ability to solve this task, we see that the two largest and most richly trained models, RoBERTa-L and ELECTRA-L, perform best. In line with our expectations, we see that most models score lower on resolving antecedents in incongruent contexts and higher in congruent ones. This is true for the four bidirectional PLMs, which also exhibited higher IC bias in Experiment 1. The gap is largest for ELECTRA and still substantial for RoBERTa-L and ELECTRA-L. The unidirectional models, on the other hand, show a noisier behavior, with a relatively small gap between the three types of stimuli, and an inconsistent ranking between them.

6.2 Discussion

As IC bias contributes to the construction of causal links in narrative and as such aids language comprehension (Trabasso and Sperry, 1985), it is desirable that PLMs capture and use this signal coming from the lexical semantics of interpersonal verbs. In Experiment 1, we found that all PLMs studied show a medium to high correlation with human IC bias scores on the level of representations, with bidirectional ones doing so on the level of predictions, too. While IC bias does contribute to language comprehension, it also has the unfortunate effect of interference in the context of incongruent EC. In this respect, models with higher IC bias awareness, i.e. the bidirectional PLMs in our study, suffer a greater drop in performance. Meanwhile, the unidirectional PLMs studied, which show little awareness of IC bias in a context free of EC, also show no interference from it when resolving referents in the presence of EC. Paradoxically, the superior performance of bidirectional PLMs with respect to IC bias also exposes a limitation of theirs: while these models are advanced enough to use IC bias for their predictions, their interpretation of semantics is still fairly shallow. The lower-order signal coming from lexical semantics is given priority over the higher-order signal coming from the sentence-level semantics. In the experiment presented in this section, this leads to a higher error rate on resolving pronoun antecedents in incongruent contexts, with potential impact on tasks that depend on co-reference resolution, e.g. document summarization (Azzam et al., 1999), question answering (Morton, 1999; Vicedo and Ferrández, 2000), and information extraction (Zelenko et al., 2004).

7 Conclusion

From the comparison of six competitive PLMs, BERT, RoBERTa-L, GPT, GPT2-M, ELECTRA and ELECTRA-L, we conclude that PLMs can exhibit IC bias much like humans do, but that different models do so to a different degree, with bidirectional models showing moderate to strong correlation to human judgements, and unidirectional models showing only a weak correlation. This ability of some PLMs has the unfortunate effect that it makes them prone to higher error rates in contexts of incongruent IC and EC signals, where the PLMs overly rely on IC bias. This finding adds to a growing body of evidence that PLMs prioritize lexical cues over higher-order semantic cues (cf. Tenney et al., 2019). As our hypothesis is inspired by the observation that humans experience a delay in the processing of incongruent contexts (Caramazza et al., 1977), our findings point to the potential of drawing further inspiration from such psycholinguistic phenomena in studying the behaviour of language models (Ettinger, 2020). Seeing that language models show a growing potential as off-the-shelf task solvers

(Radford et al., 2019; Brown et al., 2020), studying their predictions is an important avenue for better understanding their capabilities and limitations.

8 Acknowledgements

We thank Daniel Hershcovich, Ana Valeria González, Emanuele Bugliarello, and Mareike Hartmann for feedback on the drafts of this paper. We thank Desmond Elliott, Stella Frank and Dustin Wright, and Mareike Hartmann for their help with the annotation of the newly developed stimuli.

Yova was funded by Innovation Fund Denmark, under the AutoML4CS project. Mark received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (FASTPARSE, grant agreement No 714150) and from the Centro de Investigación de Galicia (CITIC) which is funded by the Xunta de Galicia and the European Union (ERDF - Galicia 2014-2020 Program) by grant ED431G 2019/01.

References

Mostafa Abdou, Vinit Ravishankar, Maria Barrett, Yonatan Belinkov, Desmond Elliott, and Anders Søgaard. 2020. The sensitivity of language models and humans to Winograd schema perturbations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7590–7604, Online. Association for Computational Linguistics.

Saliha Azzam, Kevin Humphreys, and Robert Gaizauskas. 1999. Using coreference chains for text summarization. In *Coreference and Its Applications*.

Ashley S Bangert, Richard A Abrams, and David A Balota. 2012. Reaching for words and nonwords: Interactive effects of word frequency and stimulus quality on the characteristics of reaching movements. *Psychonomic bulletin & review*, 19(3):513–520.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Myriam Cantú Sánchez. 2011. Nominal plural marking of nonce words in child Spanish. Master's thesis, Universitat Autònoma de Barcelona.

Alfonso Caramazza, Ellen Grober, Catherine Garvey, and Jack Yates. 1977. Comprehension of anaphoric pronouns. *Journal of verbal learning and verbal behavior*, 16(5):601–609.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained BERT model and evaluation data. In *PML4DC at ICLR 2020*.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Christine Cuskley, Francesca Colaiori, Claudio Castellano, Vittorio Loreto, Martina Pugliese, and Francesca Tria. 2015. The adoption of linguistic rules in native and non-native speakers: Evidence from a wug task. *Journal of Memory and Language*, 84:205 – 223.

Forrest Davis and Marten van Schijndel. 2020. Discourse structure interacts with reference but not syntax in neural language models. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 396–407, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Pennylyn Dykstra-Pruim. 2003. L2 acquisition of German plurals: How students form them and textbooks teach them. *Die Unterrichtspraxis/Teaching German*, pages 43–55.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Evelyn C Ferstl, Alan Garnham, and Christina Manouilidou. 2011. Implicit causality bias in English: A corpus of 300 verbs. *Behavior Research Methods*, 43(1):124–135.

Alan Garnham, Matthew Traxler, Jane Oakhill, and Morton Ann Gernsbacher. 1996. The locus of implicit causality effects in comprehension. *Journal of memory and language*, 35(4):517–543.

Catherine Garvey and Alfonso Caramazza. 1974. Implicit causality in verbs. *Linguistic Inquiry*, 5(3):459–464.

Edurne Goikoetxea, Gema Pascual, and Joana Acha. 2008. Normative study of the implicit causality of 100 interpersonal verbs in Spanish. *Behavior Research Methods*, 40(3):760–772.

Yoav Goldberg. 2019. Assessing BERT's syntactic abilities. *arXiv preprint arXiv:1901.05287*.

Joshua K Hartshorne, Timothy J O'Donnell, and Joshua B Tenenbaum. 2015. The causes and consequences explicit in verbs. *Language, cognition and neuroscience*, 30(6):716–734.

Joshua K Hartshorne and Jesse Snedeker. 2013. Verb argument structure predicts implicit causality: The advantages of finer-grained semantics. *Language and Cognitive Processes*, 28(10):1474–1508.

Robert Hawkins, Takateru Yamakoshi, Thomas Griffiths, and Adele Goldberg. 2020. Investigating representations of verb bias in neural language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4653–4663, Online. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Emiel Van den Hoven and Evelyn C Ferstl. 2018. Discourse context modulates the effect of implicit causality on rementions. *Language and Cognition*, 10(4):561–594.

Jaap Jumelet and Dieuwke Hupkes. 2018. Do language models understand anything? On the ability of LSTMs to understand negative polarity items. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 222–231.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4(1):521–535.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202.

Kanishka Misra, Allyson Ettinger, and Julia Rayz. 2020. Exploring BERT's sensitivity to lexical cues using tests from semantic priming. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4625–4635, Online. Association for Computational Linguistics.

Thomas S. Morton. 1999. Using coreference for question answering. In *Coreference and Its Applications*.

Matthew E. Peters, Mark Neumann, Luke Zettlemoyer, and Wen tau Yih. 2018. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

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

Samuel Rönnqvist, Jenna Kanerva, Tapio Salakoski, and Filip Ginter. 2019. Is multilingual BERT fluent in language generation? In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, pages 29–36, Turku, Finland. Linköping University Electronic Press.

Ieva Staliūnaitė and Ignacio Iacobacci. 2020. Compositional and lexical semantics in RoBERTa, BERT and DistilBERT: A case study on CoQA. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online. Association for Computational Linguistics.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2019. oLMpics–On what language model pre-training captures. *arXiv preprint arXiv:1912.13283*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel Bowman, Dipanjan Das, et al. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. In *7th International Conference on Learning Representations, ICLR 2019*.

Lucas Torroba Hennigen, Adina Williams, and Ryan Cotterell. 2020. Intrinsic probing through dimension selection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 197–216, Online. Association for Computational Linguistics.

Tom Trabasso and Linda L Sperry. 1985. Causal relatedness and importance of story events. *Journal of Memory and language*, 24(5):595–611.

Shiva Upadhye, Leon Bergen, and Andrew Kehler. 2020. Predicting reference: What do language models learn about discourse models? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 977–982, Online. Association for Computational Linguistics.

Jos JA Van Berkum, Arnout W Koornneef, Marte Otten, and Mante S Nieuwland. 2007. Establishing reference in language comprehension: An electrophysiological perspective. *Brain research*, 1146:158–171.

José L. Vicedo and Antonio Ferrández. 2000. Importance of pronominal anaphora resolution in question answering systems. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 555–562, Hong Kong. Association for Computational Linguistics.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

Nathaniel Weir, Adam Poliak, and Benjamin Van Durme. 2020. Probing neural language models for human tacit assumptions. *Cognitive Science*. To appear.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Eugen Zaretsky, Hans-Helge Müller, and Benjamin P Lange. 2016. No default plural marker in Modern High German. *Italian journal of linguistics*, 28(2):203–230.

Dmitry Zelenko, Chinatsu Aone, and Jason Tibbetts. 2004. Coreference resolution for information extraction. In *Proceedings of the Conference on Reference Resolution and Its Applications*, pages 24–31, Barcelona, Spain. Association for Computational Linguistics.

A Alternative task formulation

A.1 Splitting the clauses

Hartshorne and Snedeker (2013) induce a response by presenting participants with a sentence like

(5) John praised Tim because he was a *dax*.

and asking them the question *Who do you think is a dax?* In the spirit of Radford et al. (2018) and Radford et al. (2019) we considered reformulating this into a task suitable for a language model as:

(6) John praised Tim because he was a *dax*. The one who was a *dax* was ___

This formulation would have allowed us to use unidirectional PLMs in a more natural way, for next-word prediction, since the token of interest now comes at the end of the sequence; and it would have been equally suitable for bidirectional PLMs. We performed experiments with it and found that models largely scored at chance level for the stimuli containing neutral verbs, which renders the results for the congruent and incongruent stimuli invalid. The difficulty PLMs faced in solving this task could stem from the more complex inference required and/or from the border-line unnatural structure of the inputs.

A.2 Swapping the clauses

Seeing that all models show some degree of object bias, we considered an alternative task formulation, where the main clause and the subordinate clause are swapped:

(7) Because ___ was a *dax*, John praised Mary.

In this formulation, the proximity changes for the two referents, such that now the subject is closer to the pronoun of interest. This clause-swapping can only be applied to bidirectional models for the purposes of measuring context-free IC bias as described in Section 3.1. Figure 5 shows the results obtained with this task formulation. Comparing these numbers to the ones presented in Figure 2, we see that all models show a more balanced distribution of verbs across the S-bias and O-bias classes, with the two ELECTRA models closely matching human scores. This suggests that reference proximity is indeed a factor in the choice of the pronoun. Still, we see that RoBERTa-L and especially BERT remain strongly biased towards the object of the sentence, meaning that proximity is not the only factor at play.

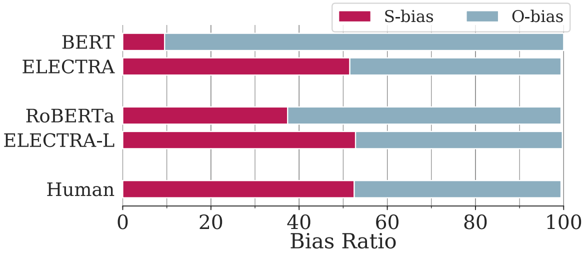


Figure 5: Ratio of S-bias verbs and O-bias verbs with swapped clauses.

Although this task formulation appears to lead to a reduced object bias across the four bidirectional models, we refrain from using it in our main experiments because it is not attested in psycholinguistic studies, i.e. it could have unforeseeable effects on human judgements of IC bias. As our main goal is to analyse the models' behavior *in relation to human behavior*, we follow closely the experimental protocol used in the available psycholinguistic studies on IC bias.

B Proper nouns

In this section we consider a seemingly minor but important consideration. Abdou et al. (2020) showed that model predictions on the Winograd Schema Challenge greatly vary with changes in the gender and identity of proper nouns used in the stimuli. We alleviate this issue by marginalizing over a range of proper nouns. To do so, we create multiple versions of the same stimulus with different proper noun combinations and use the average response of a model over all of the stimuli as an indication of the model's response to an abstract subject and object. We use 10 male names and 10 female names in 200 permutations. To ensure that the names in these lists are perceived as common names for the gender they represent, we used the models themselves to select the names, compiling a list of names unique to each model. We queried each model with the following sequences and took the top 10 names predicted:[6] *She is a woman and her name is* MASK. and *He is a man and his name is* MASK.

C Development of Materials

Neither Caramazza et al. (1977) nor Garnham et al. (1996) provide an explicit description of the procedure used to design their materials, so we extrapolate their methods by observing the materials themselves: the main goal in constructing a stimulus is to ensure that a particular ending is unambiguous (in a standard, most-likely reading) in pointing to exactly one of the two referents. The explanations always start with a verb in the past tense, e.g. *had done well*. And they are simple in the sense that they require little background knowledge. Using these observations as guidelines, we manually constructed pairs of congruent-incongruent contexts for 99 verbs, i.e. 198 stimuli in total. The materials were validated by three native English speakers and one fluent English speaker, who were asked to perform the cloze task on one stimulus from each pair and to mark ambiguous cases as such rather than making a guess at random. Eight contexts were judged as ambiguous and replaced with a better alternative, also validated in turn.

D Model implementations

We use the models as shared in the HuggingFace library (Wolf et al., 2019). In English experiments, we use bert-base-uncased, electra-large/base/small-discriminator, gpt2, and roberta-base. For German we use bert-base-german-cased, german-nlp-group/electra-base-german-uncased, dbmdz/german-gpt2, for Spanish dccuchile/bert-base-spanish-wwm-cased, and for both we use xlm-roberta-base and bert-base-multilingual-cased.

All experiments were run on a MacBook Pro, each taking between 1h (for smaller models) to 4h (for larger models).

E Probing ELECTRA

Consider the example sentence *The cake is very delicious* and a "corrupted" version of it we might present to ELECTRA: *The shoe is very delicious*. The model could give us a label sequence like {OROOO} for the latter, to indicate that shoes do not belong to the world of delicious things. And we could look at the probability distribution for the second token to quantify the strength of ELECTRA's objection to seeing this token in this position. Alternatively, ELECTRA could resolve the conflict with a label sequence like {OOOOR} to indicate that a taste-related adjective is not suitable for describing a shoe. In this case, looking at the probability distribution for the word *shoe* could be rather uninformative. This leads us to conclude that taking the average over the probability of label O for all tokens in a sequence is more informative than looking at the probability of this label for a single token.

F Experiment 1 visualisation

Figure 6 shows the data corresponding to the results described in Section 3.1. We offer this visualisation for the sake of clarity as correlation coefficients can often be misleading without the accompanying data. By and large, the visualisation of the data

[6] The mask tag and full-stop were omitted for GPT2-M. As ELECTRA cannot be used in this fashion, we instead used the generative counterpart of the model to obtain the list of names and confirmed that ELECTRA accepts them in their respective contexts (i.e. that it labels them as *original* tokens).

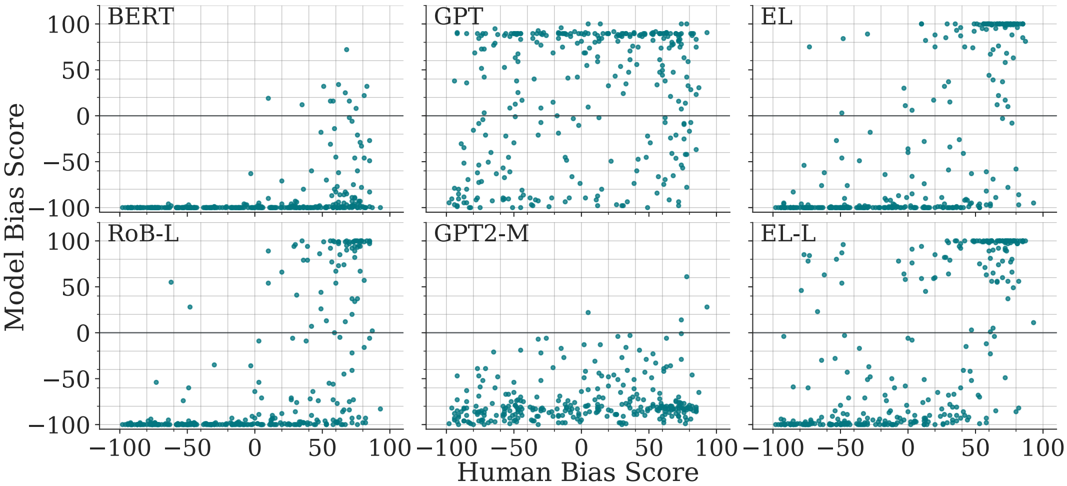


Figure 6: Model bias to human bias agreement.

corroborates the results discussed in the main body of the paper.

One observation here is that human bias scores are uniformly distributed between the two extremes (100 and -100), whereas model bias scores tend to be closer to the two endpoints. The scores obtained in the human-based study represent the responses of 100 people to the same stimulus. Deviation from the extremities here show that people differ in their judgements to some degree. The scores obtained in our study represent the responses of a single model to 200 variants of the same stimulus. In this sense, it is not surprising that the model's scores occupy the two extremities – this shows the models' consistency in judging a given verb to be subject or object biasing.

G Model representations

To obtain a single representation of any given verb from any given model, we encode a sequence like *John praised Mary* and take the representation for the first subtoken of the verb. With BERT's tokenizer, for example, the first subtoken of a verb amounts to the full verb form 57.1% of the time, and to the root of the verb 19.3%; in the remaining 22.6% the unit is 'meaningless.' These numbers vary across models, but in all cases, we are looking at contextualized embeddings, so even 'meaningless' subtokens should be a valid proxy to the verb's representation. To abstract away from the exact choice of proper nouns, we repeat this procedure for the 200 name variants and take the element-wise average over all the representations.

H Spanish and German experiments

The contexts we used for extracting proper nouns from the non-English models were *Er ist ein Mann und heißt* and *Sie ist eine Frau und heißt* for German and *Ella es una mujer y se llama* and *El es un hombre y se llama* for Spanish. For BERT DE we used *Sie heißt* and *Er heißt* instead, as these stimuli more consistently yielded names in the high ranks. Since the generator for ELECTRA DE is not publicly available, we used a different procedure to obtain the lists of personal nouns for this model: we queried WikiData for the top 100 male and female given names for people from Germany and then scored these with ELECTRA in the context shown above. We selected the 10 names for each gender that yielded highest probability of an *O* label.

For the nonce-word slot of the stimuli, we sourced nonce nouns from Dykstra-Pruim (2003) and Zaretsky et al. (2016) (German) and Cantú Sánchez (2011) (Spanish). As these lists contain less than 200 words, here sampling for the nonce-word slot in the 200 variants of a stimulus was done *with* replacement.