Language and Cognition (2024), 1–32
doi:10.1017/langcog.2024.2

ARTICLE

Does word knowledge account for the effect of world knowledge on pronoun interpretation?

Cameron R. Jones and Benjamin Bergen

Department of Cognitive Science, UC San Diego, San Diego, CA, USA
Corresponding author: Cameron Jones; Email: cameron@ucsd.edu

(Received 25 April 2023; Revised 21 December 2023; Accepted 09 January 2024)

Abstract

To what extent can statistical language knowledge account for the effects of world knowledge in language comprehension? We address this question by focusing on a core aspect of language understanding: pronoun resolution. While existing studies suggest that comprehenders use world knowledge to resolve pronouns, the distributional hypothesis and its operationalization in large language models (LLMs) provide an alternative account of how purely linguistic information could drive apparent world knowledge effects. We addressed these confounds in two experiments. In Experiment 1, we found a strong effect of world knowledge plausibility (measured using a norming study) on responses to comprehension questions that probed pronoun interpretation. In experiment 2, participants were slower to read continuations that contradicted world knowledge-consistent interpretations of a pronoun, implying that comprehenders deploy world knowledge spontaneously. Both effects persisted when controlling for the predictions of GPT-3, an LLM, suggesting that pronoun interpretation is at least partly driven by knowledge about the world and not the word. We propose two potential mechanisms by which knowledge-driven pronoun resolution occurs, based on validation- and expectation-driven discourse processes. The results suggest that while distributional information may capture some aspects of world knowledge, human comprehenders likely draw on other sources unavailable to LLMs.

Keywords: distributional baseline; distributional hypothesis; large language models; pronoun resolution; world knowledge

1. Introduction

Linguists and philosophers have long noticed distinct yet overlapping roles for 'linguistic knowledge' and 'world knowledge' in language comprehension (Frege, 1948). Linguistic knowledge refers to information that is internal to language such as

The results of experiment 1 were presented at a poster session at the 43rd Annual Meeting of the Cognitive Science Society.

grammatical agreement or semantic constraints. A sentence such as 'the professor suggested the student the idea' is ungrammatical because the verb *suggested* does not permit a dative construction without the preposition *to* (Chomsky, 1957). World knowledge, by contrast, refers to facts about the world itself which make a sentence true or false. The sentence 'Charlie Chaplin suggested the theory of relativity to Albert Einstein' is perfectly grammatical, but (as far as we know) false.

A variety of studies indicate that world knowledge has an impact on how we understand language (Warren & Dickey, 2021). We read false sentences more slowly than true ones (Garrod et al., 1994; Milburn et al., 2016), use visual information to resolve ambiguous references (Tanenhaus et al., 1995), and produce similar N400 responses to false sentences as we do to semantically implausible ones (Hagoort et al., 2004). Collectively, these kinds of results suggest that understanding language involves rapidly accessing and integrating arbitrary general knowledge about the world, which has important implications for theories of language comprehension (Barsalou, 1999; Garnham, 2001; Talmy, 2000).

In general, these studies work by manipulating whether or not a sentence is consistent with world knowledge and measuring changes in a relevant processing variable, such as reading time. If comprehenders read consistent sentences faster than inconsistent ones and relevant linguistic factors have been controlled for, we can infer that the difference in reading time must be caused by the comprehender's sensitivity to world knowledge itself. While experimenters generally control for traditional linguistic confounds such as word length and frequency, an important confound that has rarely been controlled for is the *distributional likelihood* of the expression. Words are distributed non-randomly in language and some sequences of words appear more frequently than others. In particular, because language describes the world, scenarios that are plausible in the world are also more likely to produce probable sequences of words. A growing body of work shows that comprehenders are sensitive to the distributional likelihood of expressions, above and beyond the lexical frequency of individual words (Arnon & Snider, 2010; Goodkind & Bicknell, 2018; Michaelov et al., 2022).

Until recently, state-of-the-art language models were underpowered to accurately quantify distributional likelihood in experimental stimuli (Jurafsky & Martin, 2014). However, rapid improvement in computational resources and architecture (Vaswani et al., 2017) has led to large language models (LLMs), which use neural networks to generate probability distributions over word sequences (Radford et al., 2019). LLMs serve as helpful *baselines* to measure the extent to which variance in a given phenomenon can be accounted for by distributional likelihood. They learn purely from statistical patterns in language and have no access to other innate, sensory, memory, or reasoning resources that might underlie more traditional conceptions of world knowledge (Frege, 1948; Johnson-Laird, 1989). Thus, if an LLM can account for experimental effects that have been attributed to world knowledge, it suggests that distributional information is sufficient in principle to explain the effect in humans, and undermines the claim that world knowledge is necessary to explain that effect.

In the present work, we focus on the role of world knowledge in a specific linguistic phenomenon: ambiguous pronoun resolution. The phenomenon is particularly useful as it allows us to examine the effects that world knowledge can have on a comprehender's *interpretation* of a sentence. While other paradigms show that world knowledge violations can lead to processing difficulty, this does not imply that they influence the eventual product of the comprehension process (Ferreira & Yang,

https://doi.org/10.1017/langcog.2024.2 Published online by Cambridge University Press

2019). In the case of pronominal ambiguities, however, world knowledge could fundamentally alter the propositional meaning of a sentence – the comprehender's understanding of who did what to whom. One's response to the question *Can you throw an egg at a concrete floor without cracking it?* will differ depending on how one resolves the ambiguous pronoun, *it*. This not only highlights the importance of explaining these ambiguities, it also makes them easier to study. Differing pronoun interpretations can produce discrete and radically different understandings of the sentence, often more cleanly than other types of ambiguity like polysemy.

In two experiments, we use LLMs as a *distributional baseline* (DeLong et al., 2023; Jones et al., 2022) to test whether the effects of world knowledge on interpretation can be explained by distributional linguistic information. If LLMs are able to account for knowledge effects, it would suggest that human comprehenders could also, in principle, use distributional information to resolve pronouns. This would undermine claims that non-linguistic general world knowledge is necessary for human language processing. In contrast, if world knowledge has an effect over and above distributional information, it would suggest that human comprehenders are using resources that are not available to the model when resolving pronouns, such as sensory information, embodied cognition, or general reasoning processes. This, in turn, would imply an up-front limit on the capabilities of text-only LLMs and suggest that non-linguistic information is a necessary component of human language comprehension.

In Section 1.1, we briefly survey theories of pronoun interpretation, focusing on evidence for the role of world knowledge. In Section 1.2, we discuss theoretical and empirical support for the idea that distributional information could influence human language comprehension and discuss the ways in which LLMs could be used to measure this. In Section 1.3, we briefly outline the two experiments and how their results relate to the research question.

1.1. Theories of pronoun interpretation

Words alone often fail to convey intended meanings. A reader of (1), for instance, might understand that either the baseball or the bat broke, due to the ambiguity of the pronoun *it*.

(1) When the baseball collided with the bat, it broke.

A variety of linguistic features have been found to influence ambiguous pronoun resolution. Comprehenders prefer to resolve pronouns to the subject of the previous clause (Crawley et al., 1990) or to a noun phrase that is in the same grammatical case as the pronoun (grammatical parallelism; Chambers & Smyth, 1998). Other linguistic factors, such as the semantic class of verbs, have also been found to influence pronoun resolution, including the *implicit causality* of verbs (Garvey & Caramazza, 1974). Although some researchers interpret implicit causality effects as resulting from knowledge about the typical causes of events (Pickering & Majid, 2007; Van den Hoven & Ferstl, 2018), others argue that they result from purely linguistic knowledge about verbs (Hartshorne, 2014). Finally, some pragmatic features, such as the coherence relations between sentences, have been found to alter pronoun interpretation. In a sentence completion task, Kehler and Rohde (2013) found that continuations of

https://doi.org/10.1017/langcog.2024.2 Published online by Cambridge University Press

the prompt *John passed the comic to Bill. He …* were more likely to interpret *he* as referring to John if the continuation elaborated on the first sentence, but to Bill if the continuation described a subsequent event. While the process of inferring a coherence relation between clauses might itself rely on non-linguistic world knowledge (Kehler et al., 2008), there are other cases in which surface features such as conjunctions or grammatical structure can influence coherence relations, which in turn can influence pronoun resolution.

The idea that linguistic features will govern pronoun resolution is intuitively appealing. These features are explicitly available to both producer and comprehender, minimizing the potential for miscommunication. They are also easily accessible. Memory-based models of discourse comprehension, such as the minimalist hypothesis (McKoon & Ratcliff, 1992, 2015), argue that comprehenders should only make expensive knowledge-driven inferences when they are necessary to maintain the local coherence of the text. Thus comprehenders should make use of structural features to resolve pronouns where this does not lead to incoherence.

However, in some cases, these linguistic features fail to account for our intuitions about how pronouns should be resolved. In (1), for instance, the grammatical subjecthood and parallelism biases, as well as surface features suggesting an occasion coherence relation between the clauses, all favour the subject of the previous clause (*the baseball*) as the antecedent of *it*. A reader who is familiar with baseballs and bats, however, might know that the bat is more likely to break in this case, and have an intuition that the pronoun should be resolved to the object (*the bat*). These kinds of cases have motivated researchers to posit that comprehenders can access and deploy arbitrary general knowledge during sentence parsing in order to rapidly determine which of the possible interpretations of the sentence is most plausible (Graesser et al., 1994; Hobbs, 1979; Sanford & Garrod, 1998).

There is a wide range of theoretical and empirical support for the idea that world knowledge can have this kind of influence. Constructivist theories of discourse processing argue that comprehenders routinely deploy their world knowledge to form a coherent understanding of the described situation (Graesser et al., 1994; Sanford & Garrod, 1998), and that pronouns are inevitably resolved as a by-product of this process (Garnham, 2001; Hobbs, 1979). Related psycholinguistic research shows that world knowledge can interact with other pragmatic phenomena, such as scalar implicature and the informativity of labels (Degen et al., 2015, 2019). Although many theoretical accounts of knowledge-driven pronoun resolution do not specify detailed mechanisms, some more general models of world knowledge influence provide promising candidate mechanisms for the phenomenon. One type of mechanism proposes that a comprehender's initial interpretation of a sentence is *validated* against world knowledge (O'Brien & Cook, 2016), and will be rejected or revised if an inconsistency is discovered. For example, a reader of (1) might initially use structural cues to interpret *it* as referring to *the baseball.* Upon validating this inference, the reader would recognize the inconsistency with world knowledge and revise their interpretation of *it* as referring to *the bat.* Alternatively, world knowledge might influence *expectations* about how the text will unfold before an initial interpretation has been selected (Sanford & Garrod, 1998; Venhuizen et al., 2019). In our example, a comprehender may increase their calculated probability of *broke(bat)* even before they encounter the ambiguous pronoun. Although linguistic bias will later encourage the comprehender to resolve *it* to the subject of the previous clause (*the baseball*), this might not overcome the prior, knowledge-driven bias toward *the bat.*

https://doi.org/10.1017/langcog.2024.2 Published online by Cambridge University Press

Several empirical studies provide support for knowledge-driven pronoun resolution specifically. Marlsen-Wilson and colleagues (Tyler & Marslen-Wilson, 1982a, 1982b) used contexts such as (2)–(3) to test whether participants were faster to name a congruous (*her*) or incongruous (*him*) completion following (4).

(2) As Philip was walking back from the shop, he saw an old woman trip and fall flat on her face.

(3) a. He only hesitated for a moment.
b. She seemed unable to get up again.

(4) a. Philip ran towards …
b. He ran towards …
c. Running towards …

While participants can use explicit name and gender information in (4a) and (4b) to resolve the subject to Philip, participants who heard (4c) must make the inference that the old woman was unable to run and hence is unlikely to be the subject of the clause. Nevertheless, participants showed a similar-sized delay for incongruous vs congruous probes in all three conditions. This suggests that knowledge-driven inferences can be used to resolve ambiguous references even in the absence of linguistic cues.

In a pilot study, Gordon and Scearce (1995) found that pronoun interpretation is influenced by modulating the verb in sentences like *Bill wanted John to look over some important papers… Unfortunately he never [sent/received] them.* More recently, Bender (2015) established a human baseline for the Winograd Schema Challenge: an artificial intelligence (AI) benchmark consisting of pronoun resolution problems designed to require world knowledge. Given a pair of sentences such as (5), comprehenders tended to resolve the pronoun *she* to *Ann* in (5a), but to *Mary* in (5b).

(5) a. Ann asked Mary what time the library closes, because she had forgotten.
b. Ann asked Mary what time the library closes, but she had forgotten.

While the test is used to evaluate AI models under the assumption that the knowledge-consistent answer is correct, the human baseline of 92% provided by Bender (2015) establishes empirically that human comprehenders' responses conform to the test designers' intuitions in tending to be sensitive to the plausibility of interpretations.

Although these results are consistent with the hypothesis that world knowledge influences pronoun resolution, they are also open to alternative interpretations that cannot be ruled out based on the design of the studies. First, these studies did not measure or control for other factors known to influence pronoun resolution, including the implicit causality of verbs (Garvey & Caramazza, 1974; Hartshorne, 2014) and conjunctions that alter coherence relations (Kehler & Rohde, 2013, e.g., (5)). Effects that have been attributed to world knowledge could therefore be caused by uncontrolled variance in these other factors, just as selectional restrictions and co-occurrence statistics have been found to account for world knowledge effects in other domains (Warren & Dickey, 2021; Willits et al., 2015).

https://doi.org/10.1017/langcog.2024.2 Published online by Cambridge University Press

Second, these studies do not provide any independent measure of world knowledge plausibility. The experimenter, relying on their intuition to label one antecedent as more plausible, might inadvertently be influenced by pragmatic and lexical information that was not controlled for. Third, methods in existing studies (explicit comprehension questions and cross-modal probing) could induce unnatural task demands on comprehenders, which might encourage them to deploy world knowledge more readily than they would in a more naturalistic language comprehension scenario (Ferreira & Patson, 2007). Even theories that propose a limited role for world knowledge in language comprehension acknowledge that strong task-specific incentives can motivate strategic knowledge-driven inferences (McKoon & Ratcliff, 2015). This weakens how informative existing evidence is on the stronger claim that world knowledge is deployed automatically in the course of understanding language. Finally, existing work does not control for the *distributional confound*: the possibility that distributional cues, learnable from co-occurrence statistics in language, could explain the proposed effect. We turn to this account in more detail in the next section.

1.2. Distributional information

In addition to generalizable linguistic features that influence ambiguity resolution, the rich signal of natural language provides a panoply of subtler cues. Some sequences of words appear more frequently than others and comprehenders might use their implicit knowledge of these patterns to select interpretations that are more statistically likely. The way that words are distributed in language implicitly encodes information about the world. If baseball bats are more likely to break than baseball are, then the word *breaks* might be more likely to follow *bat* than *baseball.* Even in cases where the exact sequence has never been observed before, a distributional learner can learn that *bat breaks* is more likely based on other similar contexts in which *bat* and *break* are used (Firth, 1957; Mikolov et al., 2013). A comprehender could use this statistical knowledge to resolve *it* in (1) to *the bat* by asking which of the two noun phrases is more likely to appear in the context that surrounds the ambiguous pronoun.

Although such distributional accounts of language understanding are not new (Firth, 1957; Harris, 1954, see Lenci, 2018 for discussion), the recent success of large language models has created renewed interest in these theories. Language models learn to assign probabilities to word sequences based on statistical patterns in the way that words are distributed in language. While early n-gram models simply learned transition probabilities between one sequence of words and the next, modern language models use neural networks to represent words in a multidimensional meaning space, allowing them to generalise to sequences they have never observed before (Jurafsky & Martin, 2014). Additionally, they contain attention mechanisms that allow them to relate words in the input stream to one another and represent each word differently depending on its context (Vaswani et al., 2017). Modern *large* language models are neural language models with billions of parameters trained on corpora of hundreds of billions of words or more. Some LLMs are additionally fine-tuned using reinforcement learning from human feedback (RLHF), to make their responses to input prompts safer and more useful for downstream tasks (Ouyang et al., 2022).

https://doi.org/10.1017/langcog.2024.2 Published online by Cambridge University Press

Not only do LLMs provide an explicit computational operationalization of the distributional hypothesis, but a spate of recent work shows that they are predictive of a number of human behavioural measurements, lending credence to the idea that distributional information might be sufficient to explain some aspects of human language comprehension. LLMs accurately predict a variety of measures including word relatedness judgements (Li & Joanisse, 2021; Trott & Bergen, 2021), visual similarity ratings (Lewis et al., 2019), category-membership judgements (Lenci, 2018), N400 amplitude (Michaelov et al., 2022) and reading time (Goodkind & Bicknell, 2018). Schrimpf et al. (2021) find that transformer-based LLMs predict nearly 100% of explainable variance in neural responses to sentences (fMRI and ECoG) and suggest that LLMs 'serve as viable hypotheses for how predictive language processing is implemented in human neural tissue' (p. 8).

Even in cases where we might expect world knowledge and contextual reasoning to be crucial, LLMs show an uncanny ability to mimic human response patterns. Nieuwland and Van Berkum (2007) show that human comprehenders show a large N400 response to implausible sentences such as 'The peanut was in love', except when they are preceded by a motivating context (e.g. a story about an animate peanut meeting an almond). The typical explanation of such a result is that comprehenders can use contextual information and world knowledge to process unlikely and otherwise implausible sentences. However, Michaelov et al. (2023) find that distributional models replicate the human effect, preferring the animate critical sentence to an inanimate control sentence when given the motivating story as context. This suggests that a sufficiently sensitive distributional learner can recognize that even a very globally unlikely sequence can become probable in the correct context.

To the extent that LLMs can predict human responses, it suggests that distributional information is *sufficient* to generate these responses. Although human comprehenders could still be using alternative mechanisms to reach the same results, evidence for the sufficiency of distributional information undermines claims that other resources – such as innate capacities, sensory input, or world knowledge – are *necessary* to produce the relevant behaviour. This matters because existing evidence for world knowledge influence is implicitly based on the assumption that – known linguistic factors having been controlled for – differences in responses between conditions must be attributable to non-linguistic world knowledge. A distributional language learner, however, might infer that agents who are described as *old* or have previously been the subject of *fall* are unlikely to later be the subject of the verb *to run*. Such an agent might assign a much lower probability to incongruous completion of (4), which could explain the observed reading time effect in humans (Marslen-Wilson et al., 1993).

While previous work (Kehler et al., 2004) found that predicate-argument frequency statistics did not improve the accuracy of a morphology-based pronoun resolution system, the size and complexity of modern LLMs might allow them to exploit subtler and more nuanced statistical relationships. Winograd Schemas were initially very challenging for computational models due to the deep and complex knowledge apparently required to solve them correctly. Recent advances, however, have allowed LLMs to perform as well as humans at this challenge (Kocijan et al., 2019, 2023; Sakaguchi et al., 2020). If computational models are able to resolve these ambiguous pronouns with access only to distributional information, additional evidence would be required to make a case that human comprehenders are drawing on non-linguistic world knowledge directly, rather than using the same distributional information available to language models.

https://doi.org/10.1017/langcog.2024.2 Published online by Cambridge University Press

1.3. The present study

We present two experiments designed to control for potential confounds in existing work in order to provide a more robust estimate of the influence of non-linguistic world knowledge on pronoun resolution. We

1. develop a set of stimuli similar to (1), varying the plausibility of different ambiguous pronoun interpretations while holding linguistic factors constant;
2. norm stimuli for their degree of linguistic and world knowledge bias;
3. measure the distributional likelihood of different pronoun interpretations in our stimuli using GPT-3, an LLM;
4. explicitly probe how comprehenders resolve ambiguous pronouns using comprehension questions (experiment 1);
5. measure spontaneous pronoun resolution in the absence of explicit task demands using a self-paced reading paradigm (experiment 2);
6. predict responses in each experiment using the world knowledge bias norms, controlling for the influence of linguistic bias and distributional likelihood.

We are interested in three distinct questions, each of which has different implications for the theories discussed above. First, do we see a significant effect of world knowledge bias on pronoun resolution decisions after controlling for *linguistic bias*? Accounts that explain pronoun resolution decisions on the basis of syntactic factors (Chambers & Smyth, 1998; Crawley et al., 1990) or lexical semantics (Hartshorne, 2014) do not predict a marginal effect of world knowledge as the predictive features in these theories have been held constant across conditions in our experiments. Although these theories do not claim that structural features exhaustively determine resolution decisions, a marginal effect of world knowledge would point to a systematic way in which these theories collectively fail to predict pronoun interpretation. Empirical work suggests that pragmatic biases such as scalar implicature can attenuate potential world knowledge effects (Degen et al., 2015), and so we might expect to see a similar attenuation for pronoun interpretation where informative structural cues are available.

Second, does this effect of world knowledge persist when controlling for the *distributional likelihood* of interpretations? If LLM predictions are sufficient to explain away world knowledge effects, it would undermine the claim that humans must be using non-linguistic world knowledge to resolve these ambiguities and raise the possibility that humans could also be exploiting distributional statistics (Michaelov et al., 2022; Schrimpf et al., 2021). In contrast, however, if world knowledge continues to have an independent effect on pronoun interpretation, it will provide robust evidence that non-linguistic world knowledge influences comprehenders' interpretation in a way that cannot be captured by current state-of-the-art distributional models, and suggest a way in which these models may need to be augmented in the future if they are to achieve human-like understanding of language.

Finally, do the effects of world knowledge persist in a self-paced reading paradigm without cues to resolve the pronoun (experiment 2)? Theories which posit that expensive knowledge-driven inferences are only made strategically in response to a break in coherence (McKoon & Ratcliff, 1986, 2015) might predict an effect of world knowledge in experiment 1 (where comprehenders are encouraged to deliberate on their interpretation by a comprehension question), but not in experiment 2 (where comprehenders could form an alternative coherent interpretation of the passage

https://doi.org/10.1017/langcog.2024.2 Published online by Cambridge University Press

without drawing on world knowledge. A marginal effect of world knowledge in experiment 2 would suggest that non-linguistic world knowledge is deployed spontaneously, even in the absence of specific task demands or cues (Garnham, 2001; Hobbs, 1979; O'Brien & Cook, 2016; Sanford & Garrod, 1998; Venhuizen et al., 2019).

2. Experiment 1

In experiment 1, we tested whether knowledge about the plausibility of physical events would influence pronoun resolution. Participants in the main experiment read sentences such as (6a) or (6b) and then responded to comprehension questions that indirectly probed their interpretation of the pronoun (e.g. *What broke?*).

(6) a. When the vase fell on the rock, **it** broke.
b. When the rock fell on the vase, **it** broke.

In each sentence, we refer to the first noun phrase (e.g. *the vase* in (6a)) as NP1 and the second noun phrase (e.g. *the rock* in (6a)) as NP2. Collectively we refer to these noun phrases (NPs) as the candidate antecedents. The only difference between the two versions of the sentence is that the order of the NPs is swapped. We are interested in the proportion of participants who resolve the pronoun to NP2 in each case.

We held linguistic factors discussed in Section 1.1.1 constant across the versions of each item. In both cases, NP1 is the subject of the previous clause, meaning it is favoured by both the subject assignment and grammatical parallelism biases. The lexical semantics of the two sentences are identical except for the fact that the positions of *rock* and *vase* are reversed, so any semantically-induced subject or object biases should be identical between the sentences. Finally, the surface features influencing coherence relation between the clauses were identical across versions. We refer to these non-world knowledge factors as *linguistic bias*. Orthogonally, to the extent that a comprehender's commonsense knowledge about the physical world influences their pronoun resolution decision, each sentence also has some latent world knowledge bias. For example, if the participant knows that vases are more fragile than rocks then they might be biased toward NP1 in (6a), but toward NP2 in (6b). We refer to this knowledge-driven influence as the *world knowledge bias*.

In order to independently measure the strength of the linguistic and world knowledge biases we ran two norming studies using modified versions of the stimuli. For the linguistic bias norming study, we replaced the NPs in each experimental item with two NPs deemed equally likely to participate in the critical event, such as in (7).

(7) a. When the purple vase fell on the green vase, **it** broke.
b. When the green vase fell on the purple vase, **it** broke.

There is no commonsense reason why a purple vase should be more or less likely to break than a green vase, so participants' pronoun resolution decisions should be wholly driven by linguistic factors (such as grammatical role). We confirmed this by checking that there were no large differences between responses to version a) and version b) (i.e. the linguistic bias for each version should be the same). We operationalized the linguistic bias for an item as the proportion of participants who responded with NP2 in the linguistic bias norming study.

https://doi.org/10.1017/langcog.2024.2 Published online by Cambridge University Press

Second, in order to measure the world knowledge bias for each item, we reframed the pronoun resolution problem as an explicit hypothetical reasoning question:

(8) a. If a vase fell on a rock, which would be more likely to break?
b. If a rock fell on a vase, which would be more likely to break?

Here, linguistic factors ought to have no influence as participants are explicitly encouraged to reason about the physical situation using their knowledge about the world. Again we can confirm this by checking that the bias is inverted between versions (if the bias for version (a) is 0.1, the bias for version (b) ought to be around 0.9).

If participants are guided purely by the surface cues discussed above, then there should be no difference in the proportion of participants who respond with NP2 between (6a) and (6b). Furthermore, their responses should be predicted by the linguistic bias values elicited in the norming study. In contrast, if comprehenders are deploying physical world knowledge in order to select the most plausible interpretation, they will select the same antecedents as the participants who were asked explicit reasoning questions in the world knowledge norming study. That is, there will be a positive effect of world knowledge bias on pronoun interpretation, even when controlling for the influence of linguistic bias.

We used LLMs as a distributional baseline to control for the possibility that effects could be driven by uncontrolled variance in the probability of word sequences. We included LLM responses for each item as a predictor in our regression model and tested whether world knowledge explained independent variance, just as one might control for word frequency in a lexical decision task. To the extent that participants are using distributional knowledge to resolve pronouns, probabilities assigned to sequences by an LLM should explain variance in human responses. Yet if humans are still using non-linguistic world knowledge – not learnable from language alone – to resolve pronouns, then we expect that world knowledge bias will explain additional variance even when controlling for the LLM responses.

2.1. Norming studies

2.1.1. Method

2.1.1.1. Participants. All research was approved by the UC San Diego Institutional Review Board. We recruited 35 native English-speaking undergraduate students from the UC San Diego Psychology Department subject pool, who provided informed consent using a button press and received course credit as compensation for their time. All participants successfully answered $\geq 2/3$ attention check trials. We excluded 1 participant who indicated they were not a native English speaker and 1 participant who took over 1 hour to complete the experiment. We excluded 43 trials where the response time was <500 ms (indicating guessing), and 55 trials where the response time (offset by 191 ms per syllable of question length) was >10 s (indicating inattention or excessive deliberation). We used 191 ms/syllable based on an estimate of the mean reading speed for English (Trauzettel-Klosinski et al., 2012). We retained 892 trials (463 world knowledge, 429 linguistic) from 33 participants (17 world knowledge, 16 linguistic; 23 female, 8 male, 2 non-binary; mean age = 20.3, $SD = 1.8$). The world knowledge norming study lasted 7.3 min on average ($SD = 2.2$), while the linguistic

https://doi.org/10.1017/langcog.2024.2 Published online by Cambridge University Press

norming lasted 20.6 min on average ($SD = 6.4$). The difference in duration was due to the inclusion of filler items in the linguistic norming study.

2.1.1.2. Materials. We created two alternate versions of each of the critical items from the main experiment (see Section 2.2). To elicit linguistic bias norms, we replaced the candidate antecedents with two objects that we deemed equally physically plausible. We used either modifiers that did not alter the physical properties relevant to the plausibility of the candidate, or different objects that were similar in relevant properties. To elicit world knowledge norms we reframed the pronoun resolution problem as an explicit reasoning task (see Table 1). All materials, data, and analysis code that support these results are available on the Open Science Framework at https://osf.io/v8rjm/.

2.1.1.3. Procedure. The experiment was designed using jsPsych (De Leeuw, 2015) and hosted online. Passages were presented for 250 ms + 191 ms/syllable (Trauzettel-Klosinski et al., 2012). A question then appeared below the passage with two response options. In the world knowledge norming study, the question was presented immediately and the response options were revealed after a delay. Participants used the keyboard to indicate their responses. Participants saw two examples with instructions on how to respond in each case. The examples were counterbalanced with respect to presentation order, and (in the linguistic bias norming) did not require the use of physical inference to resolve. Participants in both norming tasks were presented with 30 critical items and 3 attention check trials. Participants saw a randomly selected version of each critical item (e.g. either (7a) or (7b)). In attention check trials, participants answered simple binary questions (e.g. 'which word contains more letters: elephant or dog'). In the linguistic norming study, 45 filler items were included in order to mask the purpose of the study from participants. Filler items were taken from other pronoun resolution studies (Bender, 2015; Crawley et al., 1990; Smyth, 1994). Filler items did not encourage physical inference and were balanced with respect to NP1/NP2 bias. The presentation order of items was also randomized. The position of response options was also randomized so that the NP1 response appeared on the right in half of the trials.

Table 1. Experiment 1 example item versions and responses

	Study	Order	Stimulus	NP2 responses (%)
1	Main experiment	A	When the vase fell on **the rock**, it broke.	12.5
2	Main experiment	B	When the rock fell on **the vase**, it broke.	95
3	Linguistic	A	When the purple vase fell on **the green vase**, it broke.	10
4	Linguistic	B	When the green vase fell on **the purple vase**, it broke.	0
5	World knowledge	A	If a vase fell on **a rock**, which would be more likely to break?	0
6	World knowledge	B	If a rock fell on **a vase**, which would be more likely to break?	100

Note: In the main experiment (rows 1–2), we measured the proportion of responses that resolved an ambiguous pronoun to the second of two noun phrases (NP2, in bold). In the linguistic norming study (rows 3–4) we replaced experimental NPs with two NPs that were similar in relevant physical characteristics, in order to measure how the linguistic structure of the sentence biased interpretation. In the world knowledge norming study (rows 5–6), we reframed the pronoun resolution problem as an explicit physical reasoning task, to measure the plausibility of interpretations.

https://doi.org/10.1017/langcog.2024.2 Published online by Cambridge University Press

2.1.1.4. Results. Responses were aggregated by item to find the proportion of NP2 responses in each norming study. Results for a single item are shown in Table 1. Items in the linguistic bias norming study elicited responses that were heavily skewed toward NP1 (see Figure 1). This is likely due to subjecthood bias (as NP1 was often the subject) and grammatical parallelism (as ambiguous pronouns were often grammatical subjects). We confirmed that differences between the NPs were not influencing decisions by calculating the mean absolute difference in the proportion of NP2 responses when the order of the NPs was reversed ($M = 0.189, SD = 0.178$). Most responses in the world knowledge norming study elicited 0% or 100% NP2 responses, indicating high agreement and reflecting the fact that reversing the order of each item effectively reverses its bias with respect to NP1/NP2-coding. We confirmed that the order of the two NPs was not influencing decisions in the world knowledge norming study by checking that the mean absolute difference in proportion of NP2 responses between item versions was close to 1 ($M = 0.900, SD = 0.160$).

2.2. Main experiment

2.2.1. Methods

2.2.1.1. Participants. Participants were recruited, excluded, and compensated in the same manner as described for the norming studies. 48 participants were recruited, and 7 were excluded (5 non-native English; 1 failed $\geq 2/3$ attention check trial; 1 with completion time > 1 h) leaving 41 (25 female, 13 male, 1 non-binary, 2 prefer not to

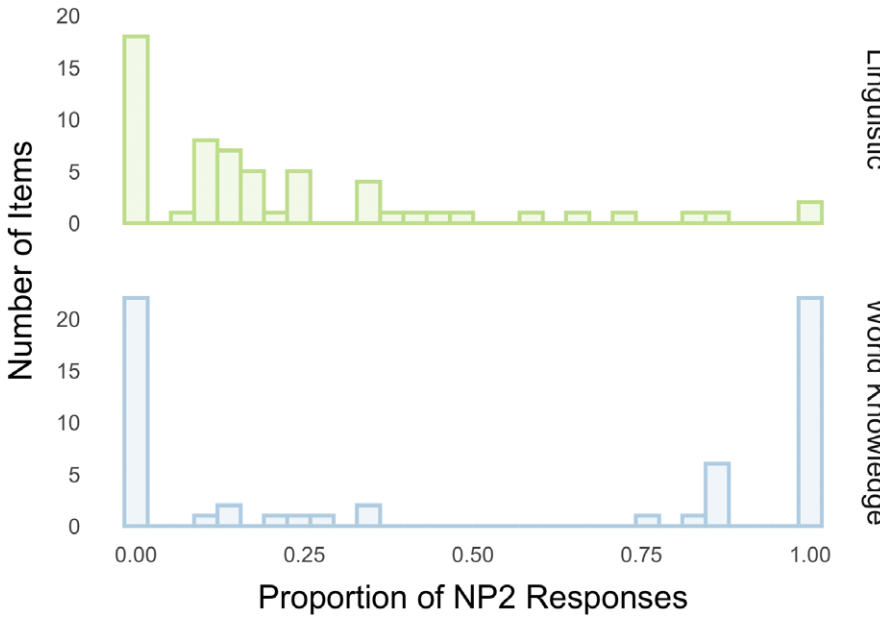

Figure 1. We used norming studies to independently measure the linguistic and world knowledge bias toward each of the noun phrases (NP1 and NP2) in our stimuli. Linguistic bias (top, green) was unimodally skewed toward NP1, likely reflecting the effects of subjecthood and grammatical parallelism biases. World knowledge bias (bottom, blue) was bimodally and symmetrically distributed, indicating high agreement and reflecting the fact that reversing the order of each item effectively reverses its bias with respect to NP1/NP2-coding.

https://doi.org/10.1017/langcog.2024.2 Published online by Cambridge University Press

say; mean age = 20.3, $SD = 2.6$). Mean completion time was 20.6 minutes ($SD = 7.4$). We excluded trials where response time was <500 ms (46) or >10 s (+191 ms/syllable, 105) leaving 1079.

2.2.1.2. Materials. Thirty critical items were designed so that each featured an introductory clause that referred to two objects (the candidates) and an ambiguous pronoun that referred back to one of the candidates in a later clause. The latter clause described a physical event in which one of the candidates was a more plausible participant than the other, such as in (6). We used a variety of situations, which would require invoking different physical properties to infer the most plausible candidate, including mass, velocity, momentum, brittleness, mass distribution, surface area, scratch hardness, indentation hardness, melting point, and flammability. We created novel stimuli to minimize the risk of dataset contamination: the possibility that LLMs have already been exposed to the stimuli in their pre-training dataset. All items were designed so that the candidates could be switched and the order of the candidates was randomized across participants, forming pairs (see Table 1, rows 5–6).

2.2.1.3. Procedure. The main experiment proceeded exactly as the linguistic bias norming study, described above (including the same instructions and filler items).

2.2.1.4. LLM analysis. We elicited predictions for each item using an LLM, GPT-3 (Brown et al., 2020). We selected GPT-3 because it is one of the best-performing LLMs that is available to the general public, and because it performs particularly well in a zero-shot setting, where it is not fine-tuned on a specific task. More specifically, we used GPT-3 *text-davinci-002*, a 175bn parameter model that has been pre-trained on more than 200bn words and additionally fine-tuned on user requests. We chose not to use later models in the GPT series because they have been additionally fine-tuned using RLHF. RLHF introduces an additional training signal beyond the likelihood of word sequences in language, making these models unsuitable for measuring how far language statistics alone can account for an effect. We accessed GPT-3 *text-davinci-002* (henceforth, GPT-3) through the OpenAI API. Following the method used for pronoun resolution problems by Brown et al. (2020), we replaced the pronoun in each stimulus with each of the candidate antecedents and elicited the sum log probability of the tokens that followed the pronoun. For (6a), this meant finding:

(9) a. $log(p(\text{broke.} \mid \text{When the rock fell on the vase, the rock}))$
b. $log(p(\text{broke.} \mid \text{When the rock fell on the vase, the vase}))$

Importantly, the model is not asked to estimate the likelihood of the candidate antecedent itself. Instead, the model's estimate of the completion of the sentence is conditioned on the pronoun being replaced by the antecedent. This allows us to measure the likelihood that the model assigns to the completion of the sequence, given that the pronoun is taken as referring to a given antecedent. This method has been found to be effective for knowledge-driven pronoun resolution in other settings, such as the Winograd Schema Challenge (Kocijan et al., 2023; Radford et al., 2019). We used the logistic function to transform the log odds ratio ((9b)–(9a)) into a probability of the model selecting NP2.

https://doi.org/10.1017/langcog.2024.2 Published online by Cambridge University Press

2.2.2. Statistical analysis

We constructed mixed-effects logistic regression models using the *lme4* package (v1.1.31; Bates et al., 2007) in R (v4.2.2; R Core Team, 2013). Regression models predicted the proportion of NP2 responses for each item version in the main experiment. We fit a maximal random effects structure (Barr et al., 2013) in order to minimize the risk of spurious explanatory power being attributed to our fixed effects. Each model contained random slopes for world knowledge bias and linguistic bias by participant, and random intercepts by participant and by item-version nested within item. We used Likelihood Ratio Tests to perform nested model comparisons that measured the predictive value of adding additional predictors to a null model with random effects only. Our full model structure was as follows:

$$\begin{aligned}&\text{response} \sim \text{world\_knowledge\_bias} + \text{linguistic\_bias} + \text{gpt3\_pNP2} + \\ &(\text{world\_knowledge\_bias} + \text{linguistic\_bias} + \text{gpt3\_pNP2} \,|\, \text{participant}) + \\ &(1 \,|\, \text{item/version}).\end{aligned}$$

2.3. Results

No significant effect of linguistic bias was detected compared to a null model with only random effects ($\chi^2(1) = 0.387$, $p = 0.534$; marginal $R^2 = 0.002$; see Figure 2). Distributional likelihood (operationalized as GPT-3 probabilities) significantly improved the fit of a model with only linguistic bias as a predictor ($\chi^2(1) = 20.1$, $p < 0.001$; marginal $R^2 = 0.121$). World knowledge had a significant effect on responses when controlling for linguistic bias only ($\chi^2(1) = 65.2$, $p < 0.001$; marginal $R^2 = 0.549$), and

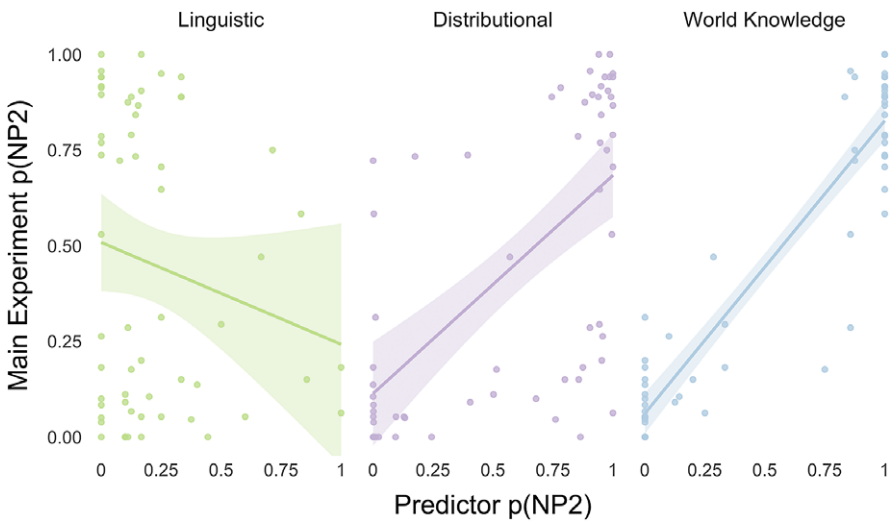

Figure 2. Left: Linguistic factors, such as grammatical role, had little influence on whether the second noun phrase (NP2) was selected as an antecedent ($r = -0.129$, $\chi^2(1) = 0.387$, $p = 0.534$). Centre: Distributional likelihood (operationalised as GPT-3 probability) was positively correlated with pronoun resolution decisions, and explained significant variance controlling for linguistic factors ($r = 0.482$, $\chi^2(1) = 20.1$, $p < 0.001$). Right: The world knowledge plausibility of NP2 positively predicted whether it is selected as an antecedent, controlling for linguistic and distributional factors ($r = 0.714$, $\chi^2(1) = 50.8$, $p < 0.001$).

https://doi.org/10.1017/langcog.2024.2 Published online by Cambridge University Press

Table 2. The full model predicting pronoun resolution decisions in experiment 1.

	Estimate	*SE*	*z* value	*p*-value
Intercept	−4.053	0.478	−8.478	< **0.001**
World knowledge	5.358	0.610	8.785	< **0.001**
Linguistic bias	0.050	0.643	0.077	0.939
Distributional likelihood	0.928	0.483	1.921	0.055

Note: There was a significant effect of world knowledge even after controlling for the other predictors.
Bold typeface indicates *p*-values < 0.05.

when controlling for both linguistic bias and distributional likelihood ($\chi^2(1) = 50.6$, $p < 0.001$; marginal $R^2 = 0.555$).

The full model showed a significant positive effect of world knowledge bias ($\beta = 5.56, p < 0.001$), and nonsignificant effects of GPT-3 predictions ($\beta = 1.38$, $p = 0.171$) and linguistic bias ($\beta = 0.028$, $p = 0.967$, see Table 2). The result shows that world knowledge bias explains additional variance in responses which is not accounted for by linguistic or distributional information. Consequently, world knowledge appears to affect interpretation in ways that cannot be explained away by existing linguistic models or the distributional knowledge account.

We performed follow-up analyses to better understand the divergence between world knowledge and distributional information. There was a fairly strong correlation between the distributional likelihood of a response and the world knowledge bias toward it ($r = 0.548$). GPT-3 predictions did not improve the fit of a model with linguistic bias and world knowledge as predictors, indicating that distributional information does not explain independent variance from these measures ($\chi^2(1) = 2.80$, $p = 0.094$). In order to test whether the world knowledge variable benefited from being less graded (and hence more decisive) than distributional likelihood, we ran a follow-up analysis with transformed GPT-3 probabilities, that had been binned into the number of unique values in the world knowledge variable (13).[1] However, the pattern of results was very similar (DL vs LB: $\chi^2(1) = 19.6$, $p < 0.001$, marginal $R^2 = 0.120$; WK vs DL + LB: $\chi^2(1) = 50.7$, $p < 0.001$, marginal $R^2 = 0.556$).

Overall GPT-3 preferred the more physically plausible antecedent on 73% of items, compared to 85% for human comprehenders. Of the 16 items where GPT-3 produced an answer that was inconsistent with physical world knowledge, 8 were paired versions from the same 4 item templates (i.e. GPT-3 produced knowledge-inconsistent answers on both version A and version B of that item, suggesting that its implicit representation of the physical world was inconsistent with that of human comprehenders). All of these items involved relatively complex physical interactions that took place over time in sealed containers (e.g. whether a lime or a can of tomatoes would be squashed in a shopping bag; a shirt or a book would be creased in a suitcase; a cardboard or steel box would be crushed in a moving van; keys or coins would create a hole in a pocket), suggesting that the model's representations may not be sufficiently fine-grained to infer the results of more involved physical interactions. The other 8 errors were caused by distinct items (i.e. GPT-3 produced a knowledge-consistent answer on the reversed counterpart version). In each case, GPT-3 predicted NP2 for both item versions. In 6/8 cases this was inconsistent with the linguistic bias as measured in the norming study. This suggests that GPT-3 was also

[1]We thank an anonymous reviewer for this suggestion.

https://doi.org/10.1017/langcog.2024.2 Published online by Cambridge University Press

making use of some structural cues (though different ones than human comprehenders) to make predictions and that in some cases these cues were strong enough to override any influence of physical plausibility.

2.4. Discussion

Participants were more likely to select an NP as the antecedent of a pronoun if the NP was judged to be a more plausible participant in the described event. In contrast, the linguistic bias of the sentence – exerted by grammatical features and measured in the linguistic norming study – did not show a significant effect on pronoun resolution decisions. Although the distributional likelihood of an interpretation (measured using GPT-3) had a significant effect on comprehenders' responses, world knowledge bias improved model fit when controlling for both linguistic and distributional information.

The results suggest that non-linguistic world knowledge does exert an influence on pronoun resolution. They also provide more robust evidence that pronoun resolution cannot be explained purely by syntactic, lexical, and discourse coherence factors (Crawley et al., 1990; Grosz et al., 1995; Hartshorne, 2014). Moreover, they suggest that while LLMs can implicitly represent *some* of the world knowledge comprehenders are using to resolve ambiguities, a large portion of the effect of world knowledge is not currently captured by these models. The result is inconsistent with the distributional hypothesis and the claims that large language models are approximating the human language comprehension process (Schrimpf et al., 2021). Instead, the effect confirms the prediction of accounts which argue that comprehenders activate relevant world knowledge during language comprehension in order to resolve ambiguities in the linguistic signal by selecting the most plausible interpretations (Hobbs, 1979; Sanford & Garrod, 1998).

There are several limitations of the study, however, which limit the generalizability of the finding. First, the passages are very short (1–2 sentences) and so they might lead participants to engage special strategies which are not representative of more ecologically typical reading behaviour of longer runs of text (van den Broek et al., 2011; Zwaan & Van Oostendorp, 1993). Second, we probe participants' pronoun resolution decisions by asking them explicit comprehension questions. This provides participants with a crucial opportunity and motivation to reason deliberatively about the plausibility of the interpretation. It may be this question-induced reasoning that leads to the deployment of world knowledge, rather than the ambiguous pronoun itself (McKoon & Ratcliff, 2015). In an attempt to address some of these concerns while replicating this result, we ran a follow-up study with several modifications: i) We embedded critical sentences within longer passages in order to lower the salience of the ambiguous pronoun, ii) We presented comprehension questions on a separate page from the passage so participants could not re-read the passage after reading the question, and iii) we included two filler comprehension questions in order to lower the salience of the critical comprehension question. The pattern of results was the same as the original experiment and world knowledge bias explained additional variance when controlling for linguistic and distributional information ($\chi^2(1) = 48.3$, $p < 0.001$; see Appendix A). However, this replication continued to provide participants with a crucial opportunity for strategic reasoning by asking a comprehension question about the critical pronoun. We addressed this limitation in experiment 2 by using self-paced reading to detect participants' spontaneous pronoun resolution decisions more indirectly.

https://doi.org/10.1017/langcog.2024.2 Published online by Cambridge University Press

3. Experiment 2

Theories of language comprehension distinguish between *strategic* and *automatic* inferences (Long & Lea, 2005; McKoon & Ratcliff, 1992). Automatic processes are fast, outside of conscious control, and insensitive to contextual factors. Strategic processes are slow, deliberate, and sensitive to the specific goals of the reader. Determining whether a process is automatic or strategic is crucial for understanding whether an observed effect is an invariant component of the language comprehension system or an artefact of specific task demands (McKoon & Ratcliff, 2015).

The results of experiment 1 could be driven by a process that automatically activates world knowledge and selects the most plausible interpretation of the pronoun: when answering comprehension questions, participants would simply recall the entity that they have encoded as the referent of the pronoun (Hobbs, 1979; Sanford & Garrod, 1998). Alternatively, the effect could result from specific features of the task that motivate strategic reasoning about physical plausibility. Specifically, participants might not perform any knowledge-driven inference during reading and only deploy world knowledge when they are presented with the comprehension question. Previous work has suggested that comprehenders do not always uniquely resolve pronouns (Greene et al., 1992), and may produce only a "Good Enough" interpretation of the text unless specific cues or task demands require them to process it more deeply (Ferreira & Patson, 2007). It could be that comprehenders would ordinarily forego expensive knowledge-driven pronoun resolution unless they are specifically incentivized to strategically deploy this process.

Fortunately, these interpretations make divergent predictions. If participants are automatically deploying world knowledge to resolve the pronoun, then the results of their inference should be available soon after reading the critical sentence and should influence how they interpret later sentences. For example, after reading (10), a comprehender might infer that *it* refers to *the vase*, and therefore that the vase is broken. This comprehender should have no difficulty in subsequently integrating the assertion in (11a), which is consistent with their current situation model. However, if (10) is instead followed by (11b), the comprehender will encounter a contradiction. The vase, which they had inferred was broken, is *still intact*.

(10) When the rock fell on the vase, it broke.

(11) a. Jennifer darted over to the shelf and saw that the *rock* was still intact on the floor.
b. Jennifer darted over to the shelf and saw that the *vase* was still intact on the floor.

Existing work demonstrates that comprehenders read more slowly when a text contradicts their current situation model (Albrecht & O'Brien, 1993; van Moort et al., 2018). Therefore, theories which claim that comprehenders are automatically deploying world knowledge predict that participants will read continuations like (11b) more slowly than control sentences that contain no inconsistency. In contrast, if comprehenders deploy world knowledge only strategically during question answering, we should observe no such slowdown. We conducted a self-paced reading study to test whether participants were slower to read continuations which contradicted more plausible pronoun interpretations.

https://doi.org/10.1017/langcog.2024.2 Published online by Cambridge University Press

3.1. Methods

3.1.1. Participants

A total of 205 participants were recruited and compensated as described in experiment 1. A larger sample was used due to an increase in the number of experimental conditions from 2 to 8. We excluded 37 participants for indicating they were not native English speakers; 14 who were inaccurate in $>50\%$ of attention check questions; 5 who took over an hour to complete the experiment (indicating inattention); 7 who indicated they did not have normal or corrected-to-normal vision; and 1 who indicated they were dyslexic, retaining 141 (88 female, 47 male, 6 non-binary; mean age = 21.6, $SD = 2.9$). Mean completion time was 19.8 minutes ($SD = 7.2$). From 4,230 trials we excluded 283, retaining 3,947. We excluded trials where passage reading times were <50 ms/syllable (91) or <350 ms/syllable (97), indicating inattention. We also excluded trials where the reading time for any recorded region was <100 ms (52), >5 s (38).

3.1.2. Materials

Thirty stimulus passage templates were designed based on the stimuli from experiment 1. Each passage contained six sections (see Figure 3). The introduction section (2–4 sentences) mentioned each candidate exactly once in the same grammatical role. Half of the passages mentioned the more plausible candidate first. The setup sentence – a buffer between the introduction and critical sentence – did not refer explicitly to either candidate. The critical sentence was identical to its respective experiment 1 stimulus and the critical spillover ensured that participants had time to make the pronoun resolution inference. It did not mention either candidate or any information that would influence interpretation of the pronoun. The continuation sentence described one of the candidates in a state that was inconsistent with it having been the antecedent of the critical pronoun and the continuation spillover was used to record delayed reading slowdowns. One comprehension question was designed for

Introduction: Jennifer was admiring the collection of marvelous objects on her bookshelf. There was a small marble bust of Socrates, a hand-painted porcelain vase, and a piece of igneous rock from Mount Etna. Suddenly, she began to feel a rumbling all around the room: it was an earthquake!

Setup: She looked on in horror as objects began to fall from shelf to shelf.

Critical: When the vase$_{NP1}$ fell on the rock$_{NP2}$, it$_{REF}$ broke.

Critical Spillover: Jennifer quickly ran for cover and after a few seconds the rumbling subsided.

Continuation: Jennifer darted over to the shelf and saw that the vase$_{CONT}$ was still intact on the floor.

Continuation Spillover: A tall lamp had also fallen over, but it was still working.

Attention Check: Jennifer had a marble bust of Aristotle.

Answer: False

Figure 3. An example passage stimulus and attention check question from experiment 2. The order of the possible antecedents is counterbalanced across participants by swapping the positions of NP1 and NP2. Continuations are made to contradict one interpretation of the pronoun by asserting that one of the NPs (here, NP1) is in a state that is inconsistent with it having been the referent of the pronoun (CONT). Unambiguous, consistent control sentences are generated by replacing the referring expression (REF) with an explicit reference to the NP not mentioned in the continuation. See Table 3 for a full list of item permutations.

https://doi.org/10.1017/langcog.2024.2 Published online by Cambridge University Press

Table 3. Experiment 2 item versions

	Ambiguity	Order	Continuation	Critical sentence	Continuation sentence
1	Ambiguous	A	NP1	When the **vase** fell on the **rock**, **it** broke.	… the **vase** was still intact …
2	Ambiguous	A	NP2	When the **vase** fell on the **rock**, **it** broke.	… the **rock** was still intact …
3	Ambiguous	B	NP1	When the **rock** fell on the **vase**, **it** broke.	… the **rock** was still intact …
4	Ambiguous	B	NP2	When the **rock** fell on the **vase**, **it** broke.	… the **vase** was still intact …
5	Unambiguous	A	NP1	When the **vase** fell on the **rock**, **the rock** broke.	… the **vase** was still intact …
6	Unambiguous	A	NP2	When the **vase** fell on the **rock**, **the vase** broke.	… the **rock** was still intact …
7	Unambiguous	B	NP1	When the **rock** fell on the **vase**, **the vase** broke.	… the **rock** was still intact …
8	Unambiguous	B	NP2	When the **rock** fell on the **vase**, **the rock** broke.	… the **vase** was still intact …

Note: Versions varied across three dimensions: whether the reference in the critical sentence was ambiguous or unambiguous; the order of the two NPs in the critical sentence; and the NP to which the continuation referred.

each passage: a statement about the passage that was either true or false and was not relevant to the critical or continuation sentences. Half of the comprehension statements were false.

We created 8 versions of each passage template by factorially varying i) the order of the two NPs, ii) whether the continuation referred to NP1 or NP2, and iii) whether the critical sentence was ambiguous or unambiguous (see Table 3). As in experiment 1, we counterbalanced the order of the two NPs to ensure linguistic and world knowledge biases were not correlated. We varied whether the continuation referred to NP1 or NP2 in order to measure the effect of contradicting a more or less plausible interpretation of the pronoun. Finally, for each critical sentence, we generated a consistent unambiguous control sentence by replacing the ambiguous pronoun with an explicit reference to whichever NP was not mentioned in the continuation sentence. We did this to control for the possibility that an effect might be caused by the continuation sentence itself, rather than the inconsistency between the continuation and the interpretation of the pronoun. For instance, imagine that comprehenders read the continuation in row 4 of Table 3 (*the vase was still intact*) more slowly than the continuation in row 3 (*the rock was still intact*). This could either be because the continuation sentence in 4 contradicts the comprehender's earlier inference that the vase is broken, or because the continuation causes slower reading *per se*. If the difference is caused by the continuation sentence itself, we should see an equivalent slowdown for row 8 vs row 7, where the critical sentence unambiguously states that *the rock broke*, ensuring there is no inconsistency. If instead the slowdown is caused by an inconsistency between the continuation sentence and the pronoun interpretation, any difference in reading time between rows 3 and 4 should not be explained by reading times for unambiguous versions.

Texts were divided into regions of 2–5 words for self-paced reading presentation. Breaking the text into smaller regions (rather than entire sentences) ensured that our measurement was sensitive to smaller or more temporary processing difficulties. Region boundaries and linebreaks were consistent across conditions. We recorded

https://doi.org/10.1017/langcog.2024.2 Published online by Cambridge University Press

reading times for the region in the continuation that contradicted one interpretation of the pronoun (e.g. */was still intact/*). We also recorded from the 3 preceding regions to measure baseline reading pace, in order to control for trial-level idiosyncrasies. Finally, we recorded the 3 regions following the critical region in order to capture delayed effects, which are common in self-paced reading studies (Just et al., 1982). We number these regions 1–7, where region 4 is the critical region that contains the potentially contradictory information.

3.1.3. Procedure

The experiment was designed using jsPsych, based on a GitHub repository provided by the Utrecht Institute of Linguistics (Duijndam, 2020), and hosted online. Participants read 30 passages, broken up into regions. Participants fixated a cross at the location of the first region and pressed the space bar to reveal each region in turn. They were instructed to then read each region at their normal reading speed. Following the moving-window paradigm, only one region was visible at any time (Just et al., 1982). All other regions were replaced with an underscore. After 1/3 of passages, participants were asked to indicate whether a statement about the passage was true or false. Participants completed two practice trials before beginning the main experiment. Participants were prevented from participating in the experiment if their screen size was less than 1,000 px × 650 px or if they were using a mobile device or tablet. The text was 25 px black Open Sans presented on a pale grey background (#f5f5f5). The order of NPs in the critical sentences and the NP to which the continuation referred were randomized within-participant. On average each participant saw 7.5 items from each of the 4 combinations of these conditions. We varied whether the critical sentence was ambiguous *between* participants in order to prevent participants from recognizing that there were two different types of stimuli and comparing them directly.

3.1.4. LLM analysis

As in experiment 1, we used an LLM to control for the possibility that comprehenders could be using distributional information to resolve pronouns. For each token in each region, we elicited from GPT-3 the surprisal, $-\log_2(p)$, of the token conditioned on all preceding tokens in the passage (including all preceding tokens in the token's own region). We then summed the surprisals of each token in the region to find the overall surprisal for the region. This measure attempts to capture the extent to which reading time can be explained by the predictability of a word sequence given the previous linguistic context.

3.2. Statistical analysis

We hypothesised that we might see an effect in any of the regions 4–7, and so we test each region separately and correct for multiple comparisons. We constructed separate linear mixed-effects models to predict reading time for each region. All reported p-values are corrected for multiple comparisons using the Holm-Bonferroni method unless otherwise stated (Holm, 1979). In a base model, we predicted log reading time for each region using the following predictors: the NP mentioned in the continuation (NP1 or NP2); the linguistic bias toward the NP mentioned in the continuation; the

https://doi.org/10.1017/langcog.2024.2 Published online by Cambridge University Press

mean log reading speed for the trial across regions 1–3 (preceding the regions of interest); and the mean reading time for that region on the unambiguous control version of each item. In the full model, we added world knowledge bias toward the NP mentioned in the continuation as a predictor. We attempted to fit a maximal random effects structure with random intercepts and slopes for each predictor by participant and random intercepts by item-version nested within the item (Barr et al., 2013). The full model did not converge so we iteratively removed random slopes until we found a random effects structure that converged for all regions (random slopes for world knowledge and linguistic bias by participant; random intercepts by participant and by item-version nested within item). We used Likelihood Ratio Tests to compare the fit of models with and without world knowledge bias as a predictor.

In order to test whether distributional information could account for the effect of world knowledge, we re-performed the above analyses, including GPT-3 surprisal as a control predictor (in both the base and full models). For each model – predicting the reading time of a given target region – we included the surprisal of the target region itself as well as the 3 preceding regions (to account for delayed reading time slowdowns in response to surprising information). For example, the region 5 model contained as predictors the surprisal for regions 2, 3, 4, and 5. The formula for the full converging model was as follows:

$$\begin{aligned}&\text{rt.log} \sim \text{world\_knowledge\_bias} + \text{linguistic\_bias} + \text{continuation\_np} + \text{groups123\_rt.log} + \\ &\text{unambiguous\_rt.log} + \text{surprisal\_n} + \text{surprisal\_n1} + \text{surprisal\_n2} + \text{surprisal\_n3} + \\ &(\text{world\_knowledge\_bias} + \text{linguistic\_bias} \,|\, \text{participant}) + (1 \,|\, \text{item/version}).\end{aligned}$$

3.3. Results

World knowledge bias significantly improved the fit of the model for region 5 – the region immediately following the critical region ($\chi^2(1) = 9.94$, $p = 0.006$) – but not for any other region (see Table 4 and Figure 4). To ensure that this result was not an artefact of our unambiguous control, we re-performed our analysis without the control predictor and again found a positive effect of world knowledge bias in region 5 ($\chi^2(1) = 9.07$, $p = 0.01$) and no effect in other regions. These results indicate that participants read continuations more slowly when they contradict the more physically plausible interpretation of the pronoun. This in turn suggests that comprehenders use world knowledge to resolve ambiguous pronouns automatically; when they encounter a continuation that contradicts their knowledge-driven pronoun interpretation, they interpret it as an inconsistency and their reading is disrupted.

Including GPT-3 surprisals in the base and full models did not change the pattern of results: there was a significant positive effect of world knowledge bias on reading time in region 5 ($\chi^2(1) = 9.87$, $p = 0.007$) and no significant effect on any other region. The full region 5 model shows no significant effects of GPT-3 surprisal for any of the recorded regions (see Table 5 and Figure 5). This suggests that the effect of world knowledge cannot be captured by distributional statistics insofar as they are learned by GPT-3.

3.4. Discussion

The results from experiment 2 indicate that world knowledge is deployed during reading to resolve ambiguous pronouns. Log reading times for region 5 in the

https://doi.org/10.1017/langcog.2024.2 Published online by Cambridge University Press

Table 4. Effects of world knowledge bias on log reading time across continuation regions

Region	Beta	χ^2	*p*-value	*p* (corrected)
4	−0.036	3.631	0.057	0.122
5	0.075	9.943	0.002	**0.006**
6	0.042	4.187	0.041	0.122
7	0.021	1.812	0.178	0.178

Note: After correcting for multiple comparisons, a positive effect of world knowledge bias was detected in region 5: the region immediately following the potentially contradictory information in the critical sentence.
Bold typeface indicates *p*-values < 0.05.

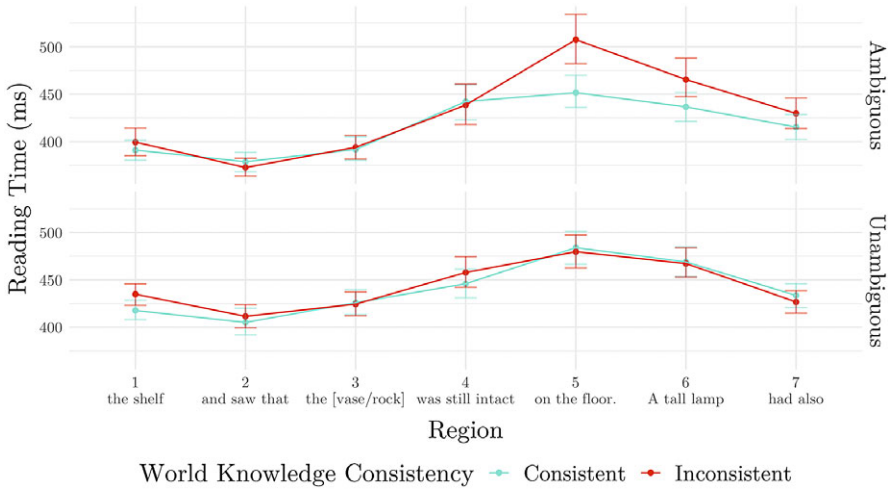

Figure 4. Mean reading time for each recorded region with 95% confidence intervals. Reading time in region 5 is 55 ms slower when the continuation is inconsistent with the more physically plausible interpretation of the ambiguous pronoun (top; $\chi^2(1) = 9.94$, $p = 0.006$). This difference is not seen when the critical sentence refers to objects unambiguously (bottom), indicating that the slowdown is due to inconsistency with the pronoun interpretation, rather than the continuation sentence itself taking longer to read.

continuation region (the region immediately following the potentially contradictory information) were positively correlated with the world knowledge bias toward the contradicted interpretation. For instance, if the critical sentence was *When the rock fell on the vase, it broke*, participants were slower to read a continuation that stated *the vase was still intact* than one that stated that *the rock was still intact.* This suggests that participants had inferred from the critical sentence that the vase was broken, and so were delayed in processing when they encountered an apparently inconsistent statement.

A potential alternative explanation is that the continuation *the vase was still intact* is simply more surprising *per se* than the continuation *the rock was still intact.* We controlled for this alternative explanation using the unambiguous consistent control, where the critical sentence explicitly referred to one NP in place of the pronoun (e.g. *When the rock fell on the vase, the rock broke*). If the continuation, *the vase was still intact*, was itself causing the slowdown in reading, we should expect to see this effect in the unambiguous control, which we do not (see Figure 4). Moreover, the world knowledge bias toward the continuation NP should not explain any additional variance on top of the control predictor, *Unambiguous log(rt)*, which it does (see

https://doi.org/10.1017/langcog.2024.2 Published online by Cambridge University Press

Table 5. The full model predicting reading times in region 5 found a significant effect of world knowledge controlling for the effect of GPT-3 surprisal

Predictor	Estimate	*SE*	*df*	*t* value	*p*-value	*p* (corrected)
Intercept	−0.986	0.511	170.484	−1.930	0.055	0.221
Physics bias	0.075	0.023	70.020	3.263	0.002	**0.007**
Structural bias	−0.045	0.033	1,849.655	−1.360	0.174	0.696
Continuation NP	−0.053	0.027	1,790.705	−1.956	0.051	0.202
Groups 1–3 log(rt)	0.781	0.031	672.699	25.081	< 0.001	< **0.001**
Unambiguous log(rt)	0.397	0.079	142.040	4.997	< 0.001	< **0.001**
Surprisal group 5	−0.000	0.001	41.304	−0.205	0.838	1.000
Surprisal group 4	0.001	0.001	44.095	0.978	0.333	1.000
Surprisal group 3	−0.001	0.001	31.701	−0.845	0.404	1.000
Surprisal group 2	−0.000	0.001	26.745	−0.298	0.768	1.000

Note: Degrees of freedom and *p*-values were calculated using the lmerTest package (v3.1.3) in R (Kuznetsova et al., 2015). Bold typeface indicates *p*-values < 0.05.

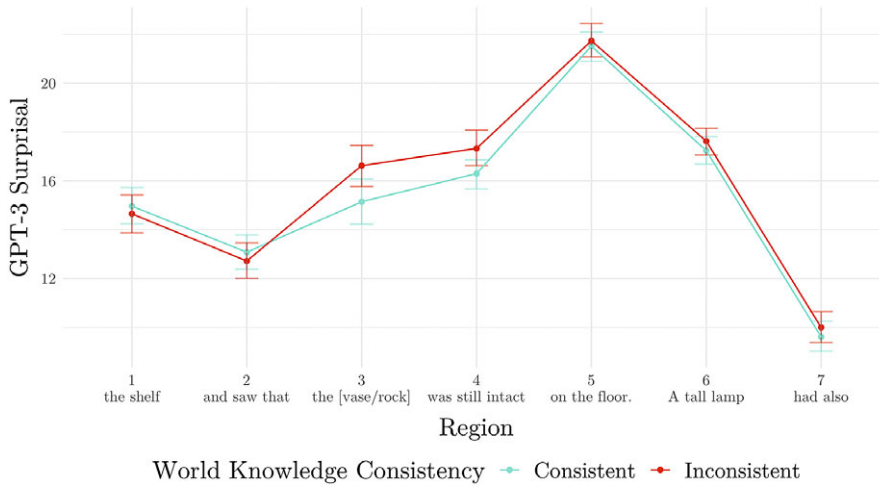

Figure 5. Mean GPT-3 surprisal for each region with 95% confidence intervals. Surprisal appears larger in regions 3 and 4 for continuations that are inconsistent vs consistent with world knowledge. However, world knowledge bias continues to have a strong positive effect on reading times for region 5 when controlling for GPT-3 surprisal across regions 2–5 ($\chi^2(1) = 9.87$, $p = 0.007$), indicating that the influence of world knowledge on human comprehenders cannot be accounted for by distributional information.

Table 5). In short, the world knowledge bias effect only occurs in the ambiguous condition, indicating that it is the result of contradicting an earlier pronoun interpretation, not of reading the continuation sentence itself.

As with experiment 1, we used an LLM as a distributional baseline to control for the possibility that participants were using information about the distribution of words in language rather than non-linguistic world knowledge to resolve ambiguous pronouns. The surprisal for each region (elicited from GPT-3) appeared to show some sensitivity to world knowledge consistency (see Figure 5). However, when we include surprisal for each region and 3 preceding regions in a baseline model, we continue to find an effect of world knowledge bias on reading time in region 5 (see Table 5). This suggests that although distributional information might capture some

https://doi.org/10.1017/langcog.2024.2 Published online by Cambridge University Press

of the physical world knowledge that humans deploy to resolve ambiguous pronouns, it is not sufficient to capture all of this variance.

Unlike experiment 1, the results of experiment 2 cannot be explained as products of a strategic reasoning process prompted by explicit comprehension questions. The results therefore indicate that participants spontaneously inferred during reading that the pronoun referred to the more physically plausible NP and hence that that NP was in some state (e.g. the vase was broken). These results confirm the predictions of the theory that comprehenders spontaneously deploy world knowledge during language comprehension to resolve ambiguous pronouns (Garnham, 2001; Garrod et al., 1994; Hobbs, 1979). The results are inconsistent with accounts that argue that such world knowledge is only deployed strategically in response to specific motivations such as comprehension questions (McKoon & Ratcliff, 2015).

4. General discussion

Together, results from these experiments provide evidence that non-linguistic world knowledge is routinely deployed to resolve referential ambiguity. Independent norms for the physical plausibility of events – established by asking a separate group of participants explicit hypothetical reasoning questions – were found to predict the majority of variance in pronoun resolution decisions (experiment 1). The effect of world knowledge bias persisted when controlling for the linguistic factors which influence pronoun resolution (again using an independent norming study), and the distributional association between candidate antecedents and the critical sentence (using an LLM, GPT-3). Finally, the world knowledge bias norms also predicted reading times for a passage continuation that was inconsistent with one interpretation of the pronoun. Specifically, the more physically plausible the contradicted interpretation was, the slower participants were to read the continuation. This last result suggests that the product of the knowledge-driven pronoun resolution inference is available to comprehenders during reading, and therefore that world knowledge is being deployed routinely and automatically.

These studies differ from previous work in important ways that alter the conclusions that can be drawn. First, they differ from results suggesting that world knowledge violations can cause processing difficulty (Hagoort et al., 2004; van Moort et al., 2018). While these studies show that world knowledge is active and available during comprehension, they do not imply that this knowledge influences the comprehender's interpretation of the sentence (Ferreira & Yang, 2019). Importantly, in our second experiment, the reading slowdown is not caused by a world knowledge violation directly: rather it is in response to an inconsistency between the continuation sentence and the prior interpretation of the pronoun. This implies that world knowledge has already been used spontaneously to alter interpretation (i.e. resolve the pronoun) before any apparent violation was discovered. Second, in contrast to previous studies suggesting that world knowledge could influence pronoun resolution (Bender, 2015; Gordon & Scearce, 1995; Marslen-Wilson et al., 1993), the results here cannot easily be explained by known linguistic confounds or distributional likelihood as these factors were measured and controlled for. This suggests that at least part of the knowledge used in pronoun resolution is not available in language alone, and must come from alternative resources such as embodiment or reasoning processes.

https://doi.org/10.1017/langcog.2024.2 Published online by Cambridge University Press

The results have implications for diverse aspects of language research, including theories of pronoun resolution, discourse comprehension, and natural language processing. First, the results imply a need to augment contemporary models of pronoun interpretation to incorporate a more explicit role for world knowledge. The linguistic features proposed by many theories – including grammatical role (Chambers & Smyth, 1998; Crawley et al., 1990; Grosz et al., 1995) and lexical semantics (Hartshorne, 2014) – were held constant across conditions in our stimuli. Under these conditions, world knowledge was found to have a strong and independent effect on interpretation. In order to accurately predict how comprehenders will resolve a given pronoun, and to provide a psycholinguistic mechanism for how an interpretation is reached, models must explicitly articulate how world knowledge influences comprehension above and beyond linguistic factors. Centering theory, for instance (Grosz et al., 1995), acknowledges world knowledge as a potential exception, however, these results suggest that it could be a crucial and constitutive part of pronoun interpretation. Similarly, Hartshorne (2014) argues that world knowledge is less plausible as a mechanism for implicit causality effects because its influence is relatively rare and peripheral in pronoun interpretation. In contrast, these results suggest that comprehenders spontaneously use plausibility to resolve ambiguous pronouns, and hence support the idea that this process could also underlie implicit causality effects. Finally, Kehler and Rohde (2018) develop a Bayesian model of pronoun interpretation based on weighting structural cues against pragmatic expectations about which referent is likely to be mentioned next. While this model neatly synthesises diverse observations about pronoun interpretation, the present results suggest a specific way in which it could be augmented: to account for the plausibility of a given interpretation, which may not be clear until after the pronoun is encountered.

Evidence for knowledge-driven pronoun resolution also has implications for discourse processing more generally. The results contrast with predictions of Minimalist accounts of language comprehension, which propose that knowledge-driven inferences are only deployed where knowledge is highly available or there is a break in local coherence (McKoon & Ratcliff, 2015). The knowledge needed to make the inferences in the present experiments was not highly available – relevant object properties were not mentioned or otherwise made salient in the text. Moreover, comprehenders would have no way of identifying a break in local coherence unless they had already activated relevant world knowledge. Minimalist accounts therefore do not predict the routine deployment and influence of world knowledge seen in these experiments. Moreover, even models that allow for world knowledge influence, such as Kintsch and Van Dijk's (1978) text comprehension model, relegate its effect to elaborating on a core interpretation that is produced before world knowledge is activated. Instead, the results presented here support a *constitutive* role for world knowledge in language comprehension. World knowledge is activated and incorporated routinely, and can influence the core propositional parsing of the sentence (Garnham, 2001; Graesser et al., 1994; Hobbs, 1979).

The spontaneous influence of world knowledge raises questions about the mechanism by which it occurs. How are comprehenders able to rapidly integrate arbitrary knowledge and assess the plausibility of different interpretations before a parse for a sentence has been selected? Two more general discourse processes, validation and expectation, provide promising candidate mechanisms. On validation accounts, comprehenders check tentative interpretations of text against their world knowledge, and reject or revise interpretations that are found to be invalid (Isberner & Richter,

https://doi.org/10.1017/langcog.2024.2 Published online by Cambridge University Press

2013; O'Brien & Cook, 2016). On expectation-driven accounts, comprehenders use world knowledge to generate predictions about how events will unfold, and use these predictions to guide comprehension (Sanford & Garrod, 1998; Venhuizen et al., 2019). While these accounts are both consistent with the present data, they are fundamentally different mechanisms and further work is needed to adjudicate between them. One approach is to vary the strength of world knowledge bias. The validation account predicts that linguistic biases will govern resolution decisions so long as the structurally preferred candidate is not *so* implausible as to be rejected. Alternatively, the expectation account predicts that world knowledge will be routinely used to direct interpretation, so that even small world knowledge biases will influence pronoun resolution decisions. Future work along these lines is needed to identify the mechanisms that support world knowledge influence in pronoun disambiguation.

The results also have theoretical and practical implications for distributional theories of language understanding. It is notable that GPT-3 predictions correlated with both world knowledge norms and pronoun resolution decisions. This suggests that the LLM has implicitly encoded some of the world knowledge information that comprehenders use to resolve ambiguous pronouns. However, the influence of world knowledge on pronoun resolution was not fully accounted for by distributional likelihood. While GPT-3 predictions explained around 12% of the variance in human responses in experiment 1, world knowledge explained around 55%, suggesting that a large portion of the influence of world knowledge is not captured by LLMs. Moreover, GPT-3 likelihood was not predictive of reading times at all in experiment 2. These results address an important confound in previous research: the possibility that apparent world knowledge effects were being driven by distributional word knowledge. More generally, the results imply that in order to understand language, human comprehenders make use of information that is not available in the linguistic signal, perhaps because perceptually obvious features are unlikely to be explicitly reported (Shwartz & Choi, 2020). This in turn implies an up-front limit on the ability of language-only models to emulate human understanding. In order to understand language in a humanlike way, models may need to be augmented with multimodal data (Zellers et al., 2021b), simulated environments (Bisk et al., 2020; Liu et al., 2022; Zellers et al., 2021a), or human norm data (Lynott et al., 2019).

The method outlined here – using LLMs as a distributional baseline – can be applied to other linguistic phenomena to understand the extent to which distributional information could account for other aspects of language understanding. Existing work in this vein suggests that distributional information can explain a large proportion of variance in brain activity (Schrimpf et al., 2021), including in response to highly contextual phenomena (Michaelov et al., 2023). Other studies suggest that models can only partially account for certain behavioural phenomena, including the influence of sense boundaries on similarity judgements (Trott & Bergen, 2023), affordances on sensibility ratings (Jones et al., 2022), and a character's knowledge state in the False Belief Task (Trott et al., 2023). Several *hybrid* theories of semantic grounding argue that comprehenders use a combination of embodied and distributional knowledge to understand language (Barsalou et al., 2008; Dove, 2011; Louwerse, 2018). The distributional baseline method and the norming studies used here allow us to quantify the extent to which different sources of information can account for specific phenomena. This could allow us to articulate more perspicuous hybrid theories and test claims about the independence or redundancy of embodied and distributional information.

https://doi.org/10.1017/langcog.2024.2 Published online by Cambridge University Press

One potential limitation of this finding is that more capable language models may be better at identifying complex statistical relationships that underlie world knowledge. Existing research suggests that as the size and training data of language models increases, so does their performance. A future, truly massive language model may be able to capture all of the variance in responses which here is explained by world knowledge. However, current language models are already psychologically implausible as models of human cognition. Children are estimated to be exposed to around 3–11 million words per year, for a total of 30–110 million words by the time they reach adult-like linguistic competence at age 10 (Hart & Risley, 1992; Hosseini et al., 2022). By contrast, GPT-3 – the model used in our analysis – has been exposed to more than 200 billion words: ~ 2000 times that of a 10 year old (Warstadt & Bowman, 2022). While larger and better-trained models may be able to tell us more about what is learnable in principle from distributional information, evidence that this is a possible mechanism for human language comprehension will need to come from more developmentally plausible models.

Finally, the results suggest that non-linguistic information and reasoning abilities exert influence on a core language comprehension process: reference assignment. Comprehenders were able to use a wide variety of physical knowledge to compare the plausibility of events while resolving pronouns. What resources underlie the rapid deployment of this physical knowledge during language comprehension? Battaglia et al. (2013) propose that humans are equipped with an intuitive physics engine (IPE), which they can use to simulate hypothetical situations and predict their outcomes. Previous research has tested this claim on non-linguistic stimuli, but future work should examine whether the IPE can also explain physical inferences during language comprehension. Similarly, Barsalou (1999) proposes that language comprehension involves relating linguistic information to multimodal perceptual symbols grounded in sensorimotor experience. Activation of embodied perceptual symbols provides an intuitively plausible hypothesis about how world knowledge can be leveraged so efficiently to influence language interpretation (Zwaan, 2016). However, more work is needed to test whether sensorimotor processes are causally involved in comprehension more generally (Ostarek & Bottini, 2021), and in knowledge-driven inference specifically.

Understanding language necessarily involves connecting words to the world around us. However, there has been much debate about whether world knowledge can influence our interpretation of what is said. These results support a tightly integrated model in which comprehenders spontaneously retrieve relevant world knowledge and assess different possible interpretations in order to select the most plausible. However, the results also raise many more questions for future research. Is world knowledge always deployed or is it only activated by some internal or external trigger? Will world knowledge always determine the interpretation of ambiguities or can other factors overwhelm its influence? Finally, do comprehenders make knowledge-driven inferences by performing formal operations on proposition-like statements, or by simulating the sensorimotor implications of different interpretations? Answering these questions will help to illustrate the mechanisms by which we make meaning from words.

Data availability statement. The materials, data, and analysis code that support the findings of this study are openly available on Open Science Framework at https://osf.io/v8rjm/.

Acknowledgements. The authors would like to thank Andy Kehler, Noortje Venhuizen, and two anonymous reviewers for thoughtful feedback on earlier versions of this paper.

Competing interest. The authors declare none.

References

Albrecht, J. E., & O'Brien, E. J. (1993). Updating a mental model: Maintaining both local and global coherence. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19(5), 1061.

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, 62(1), 67–82.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3), 255–278.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22(4), 577–660.

Barsalou, L. W., Santos, A., Simmons, W. K., & Wilson, C. D. (2008). Language and simulation in conceptual processing. In M. de Vega, A. Glenberg, & A. Graesser (Eds.), *Symbols and embodiment: Debates on meaning and cognition* (pp. 245–283). Oxford Academic.

Bates, D., Sarkar, D., Bates, M. D., & Matrix, L. (2007). The lme4 package. *R package version* 2(1), 74.

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 110(45), 18327–18332.

Bender, D. (2015). Establishing a human baseline for the Winograd Schema Challenge. In *Proceedings of the 26th Modern AI and Cognitive Science Conference*. Valparaiso University. http://cslab.valpo.edu/~mglass/maics2015papers/index.html.

Bisk, Y., Holtzman, A., Thomason, J., Andreas, J., Bengio, Y., Chai, J., Lapata, M., Lazaridou, A., May, J., Nisnevich, A., Pinto, N., & Turian, J. (2020). Experience grounds language. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)* (pp. 8718–8735). Association for Computational Linguistics.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, D., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. In *Advances in neural information processing systems* (Vol. 33, pp. 1877–1901). Curran Associates, Inc.

Chambers, C. G., & Smyth, R. (1998). Structural parallelism and discourse coherence: A test of centering theory. *Journal of Memory and Language*, 39(4), 593–608.

Chomsky, N. (1957). *Syntactic structures*. Mouton de Gruyter.

Crawley, R. A., Stevenson, R. J., & Kleinman, D. (1990). The use of heuristic strategies in the interpretation of pronouns. *Journal of Psycholinguistic Research*, 19(4), 245–264.

De Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, 47(1), 1–12.

Degen, J., Hawkins, R. D., Graf, C., Kreiss, E., & Goodman, N. D. (2019). When redundancy is useful: A Bayesian approach to 'over informative' referring expressions. *Psychological Review*, 127(4), 591.

Degen, J., Tessler, M. H., & Goodman, N. D. (2015). Wonky worlds: Listeners revise world knowledge when utterances are odd. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 37, 548–555.

DeLong, K. A., Trott, S., & Kutas, M. (2023). Offline dominance and zeugmatic similarity norms of variably ambiguous words assessed against a neural language model (BERT). *Behavior Research Methods*, 55(4), 1537–1557.

Dove, G. (2011). On the need for embodied and dis-embodied cognition. *Frontiers in Psychology*, 1, 242.

Duijndam, M. (2020). *Jspsych-spr-mw: A self paced reading with moving window experiment using jsPsych*. UiL OTS labs.

Ferreira, F., & Patson, N. D. (2007). The 'good enough' approach to language comprehension. *Language and Linguistics Compass*, 1(1–2), 71–83.

Ferreira, F., & Yang, Z. (2019). The problem of comprehension in psycholinguistics. *Discourse Processes*, 56 (7), 485–495.

Firth, J. R. (1957). *A synopsis of linguistic theory*. Blackwell.

Frege, G. (1948). Sense and reference. *The Philosophical Review*, 57(3), 209–230.

https://doi.org/10.1017/langcog.2024.2 Published online by Cambridge University Press

Garnham, A. (2001). *Mental models and the interpretation of anaphora*. Psychology Press.
Garrod, S., Freudenthal, D., & Boyle, E. (1994). The role of different types of anaphor in the on-line resolution of sentences in a discourse. *Journal of Memory and Language*, 33, 39–68.
Garvey, C., & Caramazza, A. (1974). Implicit causality in verbs. *Linguistic Inquiry*, 5(3), 459–464.
Goodkind, A., & Bicknell, K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. In A. Sayeed, C. Jacobs, T. Linzen, & M. van Schijndel (Eds.), *Proceedings of the 8th workshop on cognitive modeling and computational linguistics (CMCL 2018)* (pp. 10–18). Association for Computational Linguistics.
Gordon, P. C., & Scearce, K. A. (1995). Pronominalization and discourse coherence, discourse structure and pronoun interpretation. *Memory & Cognition*, 23(3), 313–323.
Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, 101(3), 371–396.
Greene, S. B., McKoon, G., & Ratcliff, R. (1992). Pronoun resolution and discourse models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18(2), 266.
Grosz, B. J., Joshi, A. K., & Weinstein, S. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2), 203–225.
Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. *Science*, 304, 5.
Harris, Z. S. (1954). Distributional structure. *Word*, 10(2–3), 146–162.
Hart, B., & Risley, T. R. (1992). American parenting of language-learning children: Persisting differences in family-child interactions observed in natural home environments. *Developmental Psychology*, 28(6), 1096–1105.
Hartshorne, J. K. (2014). What is implicit causality? *Language, Cognition and Neuroscience*, 29(7), 804–824.
Hobbs, J. R. (1979). Coherence and coreference. *Cognitive Science*, 3(1), 67–90.
Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2), 65–70.
Hosseini, E. A., Schrimpf, M., Zhang, Y., Bowman, S., Zaslavsky, N., & Fedorenko, E. (2022). Artificial neural network language models align neurally and behaviorally with humans even after a developmentally realistic amount of training. *bioRxiv* 2022.10.04.510681
Isberner, M.-B., & Richter, T. (2013). Can readers ignore implausibility? Evidence for nonstrategic monitoring of event-based plausibility in language comprehension. *Acta Psychologica*, 142(1), 15–22.
Johnson-Laird, P. N. (1989). *Mental models*. MIT Press.
Jones, C. R., Chang, T. A., Coulson, S., Michaelov, J. A., Trott, S., & Bergen, B. (2022). Distributional semantics still can't account for affordances. *Proceedings of the annual meeting of the Cognitive Science Society*, 44, 482–489.
Jurafsky, D., & Martin, J. H. (2014). *Speech and language processing* (Vol. 3). Prentice Hall.
Just, M. A., Carpenter, P. A., &Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology. General*, 111(2), 228–238.
Kehler, A., Appelt, D., Taylor, L., & Simma, A. (2004). The (non) utility of predicate-argument frequencies for pronoun interpretation. In *Proceedings of the human language technology conference of the North American chapter of the Association for Computational Linguistics: HLT-NAACL 2004* (pp. 289–296). Association for Computational Linguistics.
Kehler, A., Kertz, L., Rohde, H., & Elman, J. L. (2008). Coherence and coreference revisited. *Journal of Semantics*, 25(1), 1–44.
Kehler, A., & Rohde, H. (2013). A probabilistic reconciliation of coherence-driven and centering-driven theories of pronoun interpretation. *Theoretical Linguistics*, 39(1–2), 1–37.
Kehler, A., & Rohde, H. (2018). Prominence and coherence in a Bayesian theory of pronoun interpretation. *Journal of Pragmatics*, 154, 63–78.
Kintsch, W., & Van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, 85(5), 363.
Kocijan, V., Cretu, A.-M., Camburu, O.-M., Yordanov, Y., & Lukasiewicz, T. (2019). A surprisingly robust trick for Winograd Schema Challenge. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th annual meeting of the Association for Computational Linguistics* (pp. 4837–4842). Association for Computational Linguistics.

https://doi.org/10.1017/langcog.2024.2 Published online by Cambridge University Press

Kocijan, V., Davis, E., Lukasiewicz, T., Marcus, G., & Morgenstern, L. (2023). The defeat of the Winograd Schema Challenge. *Artificial Intelligence*, 103971, 1–18.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2015). Package 'lmertest'. *R package version* 2(0), 734.

Lenci, A. (2018). Distributional models of word meaning. *Annual Review of Linguistics*, 4(1), 151–171.

Lewis, M., Zettersten, M., & Lupyan, G. (2019). Distributional semantics as a source of visual knowledge. *Proceedings of the National Academy of Sciences*, 116(39), 19237–19238.

Li, J., & Joanisse, M. F. (2021). Word senses as clusters of meaning modulations: A computational model of polysemy. *Cognitive Science*, 45(4), e12955.

Liu, R., Wei, J., Gu, S. S., Wu, T.-Y., Vosoughi, S., Cui, C., Zhou, D., & Dai, A. M. (2022). *Mind's eye: Grounded language model reasoning through simulation.* Preprint, arXiv:2210.05359.

Long, D. L., & Lea, R. B. (2005). Have we been searching for meaning in all the wrong places? Defining the "search after meaning" principle in comprehension. *Discourse Processes*, 39(2–3), 279–298.

Louwerse, M. M. (2018). Knowing the meaning of a word by the linguistic and perceptual company it keeps. *Topics in Cognitive Science*, 10(3), 573–589.

Lynott, D., Connell, L., Brysbaert, M., Brand, J., & Carney, J. (2019). The Lancaster sensorimotor norms: Multidimensional measures of perceptual and action strength for 40,000 English words. *Behavior Research Methods*, 52(3), 1271–1291.

Marslen-Wilson, W., Tyler, L. K., & Koster, C. (1993). Integrative processes in utterance resolution. *Journal of Memory and Language*, 32(5), 647–666.

McKoon, G., & Ratcliff, R. (1986). Inferences about predictable events. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 12(1), 82.

McKoon, G., & Ratcliff, R. (1992). Inference during reading. *Psychological Review*, 99(3), 440.

McKoon, G., & Ratcliff, R. (2015). Cognitive theories in discourse-processing research. In A. E. Cook, E. J. O'Brien, J. Lorch, & F. Robert (Eds.), *Inferences during reading* (pp. 42–67). Cambridge University Press.

Michaelov, J. A., Coulson, S., & Bergen, B. K. (2022). So close yet so far: N400 amplitude is better predicted by distributional information than human predictability judgements. *IEEE Transactions on Cognitive and Developmental Systems*, 15, 1033–1042.

Michaelov, J. A., Coulson, S., & Bergen, B. K. (2023). *Can peanuts fall in love with distributional semantics?* Preprint, arXiv:2301.08731.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. Burges, L. Bottou, M. Welling, Z. Ghahramani & K. Q. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 26). Curran Associates, Inc..

Milburn, E., Warren, T., & Dickey, M. W. (2016). World knowledge affects prediction as quickly as selectional restrictions: Evidence from the visual world paradigm. *Language, Cognition and Neuroscience*, 31(4), 536–548.

Nieuwland, M. S., & Van Berkum, J. J. (2007). When peanuts fall in love: N400 evidence for the power of discourse. *Journal of Cognitive Neuroscience*, 18(7), 1098–1111.

O'Brien, E. J., & Cook, A. E. (2016). Coherence threshold and the continuity of processing: The RI-Val model of comprehension. *Discourse Processes*, 53(5–6), 326–338.

Ostarek, M., & Bottini, R. (2021). Towards strong inference in research on embodiment – Possibilities and limitations of causal paradigms. *Journal of Cognition*, 4(1), 5.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730–27744.

Pickering, M. J., & Majid, A. (2007). What are implicit causality and consequentiality? *Language and Cognitive Processes*, 22(5), 780–788.

R Core Team. (2013). *R: A language and environment for statistical computing*. R Core Team.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8), 9.

Sakaguchi, K., Le Bras, R., Bhagavatula, C., & Choi, Y. (2020). Winogrande: An adversarial Winograd Schema Challenge at scale. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(5), 8732–8740.

Sanford, A. J., & Garrod, S. C. (1998). The role of scenario mapping in text comprehension. *Discourse Processes*, 26(2–3), 159–190.

https://doi.org/10.1017/langcog.2024.2 Published online by Cambridge University Press

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45), e2105646118.

Shwartz, V., & Choi, Y. (2020). Do neural language models overcome reporting bias? In D. Scott, N. Bel, & C. Zong (Eds.), *Proceedings of the 28th international conference on computational linguistics* (pp. 6863–6870). International Committee on Computational Linguistics.

Smyth, R. (1994). Grammatical determinants of ambiguous pronoun resolution. *Journal of Psycholinguistic Research*, 23(3), 197–229.

Talmy, L. (2000). *Toward a cognitive semantics*. MIT Press.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217), 1632–1634.

Trauzettel-Klosinski, S., Dietz, K., & Group, I. S. (2012). Standardized assessment of reading performance: The new international reading speed texts IReST. *Investigative Ophthalmology & Visual Science*, 53(9), 5452–5461.

Trott, S., & Bergen, B. (2021). RAW-C: Relatedness of ambiguous words in context (a new lexical resource for English). In *Proceedings of the 59th annual meeting of the Association for Computational Linguistics and the 11th international joint conference on natural language processing* (Vol. 1, pp. 7077–7087). Association for Computational Linguistics.

Trott, S., & Bergen, B. (2023). Word meaning is both categorical and continuous. *Psychological Review*, 130 (5), 1239–1261.

Trott, S., Jones, C., Chang, T., Michaelov, J., & Bergen, B. (2023). Do large language models know what humans know? *Cognitive Science*, 47(7), e13309.

Tyler, L. K., & Marslen-Wilson, W. (1982a). Processing utterances in discourse contexts: Online resolution of anaphors. *Journal of Semantics*, 1(3–4), 297–314.

Tyler, L. K., & Marslen-Wilson, W. (1982b). The resolution of discourse anaphors: Some online studies. *Text – Interdisciplinary Journal for the Study of Discourse*, 2(1–3), 263–291.

van den Broek, P., Bohn-Gettler, C. M., Kendeou, P., Carlson, S., & White, M. J. (2011). When a reader meets a text: The role of standards of coherence in reading comprehension. In G. Schraw, M. T. McCrudden, & J. P. Magliano (Eds.), *Text relevance and learning from text* (pp. 123–139). IAP Information Age Publishing.

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

van Moort, M. L., Koornneef, A., & van den Broek, P.W. (2018). Validation: Knowledge- and text-based monitoring during reading. *Discourse Processes*, 55(5–6), 480–496.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (pp. 5998–6008). Curran Associates, Inc.

Venhuizen, N. J., Crocker, M. W., & Brouwer, H. (2019). Expectation-based comprehension: Modeling the interaction of world knowledge and linguistic experience. *Discourse Processes*, 56(3), 229–255.

Warren, T., & Dickey, M. W. (2021). The use of linguistic and world knowledge in language processing. *Language and Linguistics Compass*, 15(4), e12411.

Warstadt, A., & Bowman, S. R. (2022). What artificial neural networks can tell us about human language acquisition. In Lappin, S., & Bernardy, J. P. (Eds.), *Algebraic Structures in Natural Language* (pp.17–60). CRC Press

Willits, J. A., Amato, M. S., & MacDonald, M. C. (2015). Language knowledge and event knowledge in language use. *Cognitive Psychology*, 78, 1–27.

Zellers, R., Holtzman, A., Peters, M., Mottaghi, R., Kembhavi, A., Farhadi, A., & Choi, Y. (2021a). *PIGLeT: Language grounding through neuro-symbolic interaction in a 3D world.* Preprint, arXiv:2106.00188 [cs].

Zellers, R., Lu, X., Hessel, J., Yu, Y., Park, J. S., Cao, J., Farhadi, A., & Choi, Y. (2021b). MERLOT: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, 34, 23634–23651.

Zwaan, R. A. (2016). Situation models, mental simulations, and abstract concepts in discourse comprehension. *Psychonomic Bulletin & Review*, 23, 1028–1034.

Zwaan, R. A., & Van Oostendorp, H. (1993). Do readers construct spatial representations in naturalistic story comprehension? *Discourse Processes*, 16(1–2), 125–143.

https://doi.org/10.1017/langcog.2024.2 Published online by Cambridge University Press

A. Experiment 1B

A.1. Method

A.1.1. Participants

Forty-three participants were recruited and compensated in the same manner as described for experiment 1. We excluded 6 participants for indicating they were not native English speakers and 9 who were inaccurate in $\geq 20\%$ of filler questions (indicating inattention) leaving 28 (22 female, 6 male; mean age = 22.5, $\sigma = 3.7$). Mean completion time was 22.6 minutes ($\sigma = 8.0$). We excluded 61 trials where passage reading times were <50 ms/syllable or >350 ms/syllable, 14 trials where the question response time was <500 ms or >10 s. We retained 775 trials in total.

A.1.2. Materials

Thirty stimulus passages were designed based on the stimuli from experiment 1. Each passage contained four sections (introduction, setup, critical, and continuation). The introduction section (2–4 sentences) introduced the two candidates and provided an appropriate context for the event. Each candidate was mentioned exactly once in the same grammatical role. Passages were balanced with respect to whether the candidate that was favoured by physics bias was mentioned first or second in the introduction. The setup section contained one sentence that acted as a buffer between introducing the candidates and the critical sentence in order to minimise any structural effects of the order in which candidates are mentioned before the critical sentence. The setup did not refer explicitly to either candidate, but could refer to the candidates together using a generic term (such as *the objects*). The critical sentence was identical to its respective experiment 1 stimulus. The order of candidates was randomly varied among participants as in experiment 1. The continuation section (1–3 sentences) did not mention either of the candidates and was designed not to contain any information that might be more consistent with one interpretation of the ambiguous pronoun than the other. Two filler comprehension questions were designed for each passage. These probed the participants' understanding of aspects of the passage that were unrelated to the critical sentence.

A.1.3. Procedure

The experiment was designed using jsPsych and hosted online. Participants were instructed to read short passages and then answer comprehension questions about them. Each passage was presented in its entirety. Participants pressed a button when they had finished reading the passage to advance to the comprehension questions. The comprehension questions appeared one at a time. Participants indicated their chosen response using a button press, and the next question appeared immediately. After participants had completed all three comprehension questions, the next passage was presented. The order of the comprehension questions was randomized (to minimize the salience of the critical question).

A.1. Results

We constructed logistic mixed-effects models to predict responses to the critical comprehension questions using the biases elicited in the experiment 1 norming studies. All models had random slopes by participant for the effects of physics and structural bias, and random intercepts by participant and by item-version nested within item.

Including a fixed effect of linguistic bias did not improve model fit over a null model with an intercept and random effects ($\chi^2(1) = 0.230, p = 0.631$). Adding world knowledge bias as a predictor significantly improved model fit over the linguistic bias model ($\chi^2(1) = 51.6, p < 0.001$). GPT-3 predictions also improved the fit of a model with linguistic bias only ($\chi^2(1) = 21.5, p < 0.001$). World knowledge bias significantly improved the fit of a model with both GPT-3 predictions and linguistic bias ($\chi^2(1) = 33.3, p < 0.001$). These results replicate the effect of world knowledge bias on responses that was observed in experiment 1.

Cite this article: Jones, C. R., & Bergen, B. (2024). Does word knowledge account for the effect of world knowledge on pronoun interpretation? *Language and Cognition*, 1–32. https://doi.org/10.1017/langcog.2024.2