Reasoning with Polysemes: When Default Inferences Beat Contextual Information

Eugen Fischer[1] (E.Fischer@uea.ac.uk), Paul Engelhardt[2] (P.Engelhardt@uea.ac.uk), Dimitra Lazaridou-Chatzigoga[1,2,3] (D.Lazaridou-Chatzigoga@uea.ac.uk), Kate Hazel Stanton[4] (katehazelstanton@pitt.edu)

[1]School of Politics, Philosophy, Language and Communication Studies, University of East Anglia, Norwich NR4 7TJ, UK

[2]School of Psychology, University of East Anglia, Norwich NR4 7TJ, UK

[3] Department of Theoretical and Applied Linguistics, University of Cambridge, Cambridge CB3 9DP, UK

[4] Department of Philosophy & Department of Linguistics, University of Pittsburgh, Pittsburgh, PA 15260, USA

Abstract

How, and how strongly, do default comprehension inferences shape verbal reasoning? When do they lead to fallacies? We address these questions for reasoning with polysemous verbs (verbs with distinct, but related senses) and ask when their use leads to fallacies of equivocation. The 'linguistic salience bias hypothesis' specifies conditions where subordinate uses of unbalanced polysemes trigger defeasible default inferences that are supported only by the dominant sense but influence further cognition, regardless. But does this happen even where the verb is *preceded* by disambiguating context that invites subordinate interpretations *from the start*? We present three experimental-philosophy studies that address this question: We use the psycholinguistic cancellation paradigm and fixation time measurements to examine inferences from polysemous appearance verbs. We find that default inferences can beat even preceding contextual information. Beyond their psycho-linguistic interest, findings have important philosophical consequences.

Keywords: verbal reasoning; default inferences; polysemy processing; appearance verbs; linguistic salience; eye tracking; experimental philosophy

Introduction

Comprehension Inferences

Words trigger default inferences. Verbal stimuli activate concepts, i.e., bodies of information that are (i) stored in long-term memory, (ii) deployed in the exercise of higher cognitive competencies, and (iii) retrieved *by default* – i.e., automatically, even in the absence of context, as in single-word priming studies. These concepts include prototypes (Rosch, 1975; Hampton, 2006) associated with object nouns and *situation schemas* (Rumelhardt, 1978) associated with event nouns and verbs. These schemas include information about the typical features of events (e.g., instruments used), agents, and patients. Schema information supports comprehension inferences, including attributions of typical agent- and patient-properties to role fillers. It remains a subject of debate exactly how much event information (e.g., only about instruments typically used, or also about typical event locations) is activated by verbs (review: Yee et al., 2018). Even so, event knowledge clearly plays a key role in utterance interpretation (Elman & McRae, 2019) and in building mental representations of the situation described by the utterance (*situation models*; Zwaan, 2016), which provide the basis for judgments and reasoning about those situations.

Initial activation of information from situation schemas is, however, modulated by linguistic context. First, which components of a schema get activated depends upon fit with the thematic role to be filled: Sentence fragments like 'She was arrested by the ___' activate typical agents (*cop*) in post-verbal position when they leave the agent role blank (as above), but not when they leave open the patient, as in 'She arrested the ___', (Ferretti et al., 2001; *cf.* Kim et al., 2016).

Second, in incremental utterance interpretation, ever more specific schemas are activated by verbs in conjunction with subject- and object-nouns (Bicknell et al., 2010; Matsuki et al., 2011), with prepositions and syntactic constructions like verb aspect (Ferretti et al., 2007), and with simultaneous visual stimuli (Kamide et al., 2003).

Finally, also subsequent deployment of schema information is modulated by context: Initially activated information can get suppressed within 1 sec, when it conflicts with contextual information or background beliefs (Fischer & Engelhardt, 2017, *cf.* Faust & Gernsbacher, 1996). Where its suppression is complete, initially activated information will not influence further judgment and reasoning. The findings reviewed thus motivate the large question: When, and how strongly, do default comprehension inferences shape verbal reasoning?

Polysemy Processing

We address this large question in connection with a specific hypothesis about how default inferences shape reasoning with *polysemous* words, which have several distinct but related senses. On a conservative estimate, these words account for 40% of the English vocabulary (Byrd et al., 1987). We focus, more specifically on polysemous verbs.

Converging psycholinguistic evidence suggests that many irregular polysemes activate a unitary representation of semantic information that is then deployed to interpret utterances which use the word in different senses (e.g., Macgregor et al., 2015; Pylkkänen et al., 2006). This unitary representation consists in overlapping clusters of semantic features (Brocher et al., 2016; Klepousniotou et al., 2012). Different components of these unitary representations get activated in different strength by the verbal stimulus (Brocher et al., 2018): The more often the language user encounters the word in one sense, rather than another, the more strongly the features associated with that sense are activated, upon the next encounter. This means that unbalanced polysemes activate most strongly the features associated with their dominant sense.

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

Familiar 'core meaning' models suggest that verbal stimuli initially activate semantic features that are shared by all senses and underspecify the information that is contextually relevant for any specific use (Klepousniotou et al., 2008). We consider the opposite possibility, as it arises for verbs: the initially activated situation schema may *over-specify* the information relevant for a specific use. This will happen where the dominant sense of an unbalanced polyseme is associated with a rich situation schema, only some of which is relevant for interpreting a subordinate use. (E.g., 'S sees X' is associated with a complex schema that includes the features *S knows X is there* and *S knows what X* is. Only these are relevant for interpreting purely epistemic uses like 'Jack saw her point'.) In this case, interpreting the subordinate use will involve retaining as much of the initially activated information as is relevant for the specific use, and suppressing the contextually irrelevant information, in line with the *Retention/Suppression Strategy* (Giora, 2003; 2012).

Linguistic Salience Bias

Where markedly unbalanced polysemes are processed in line with this strategy, a bias is set to arise: Due to the imbalance, the situation schema associated with the dominant sense will be strongly activated (Brocher et al., 2018). The frequently co-instantiated component features of this schema will continue to exchange lateral co-activation (Hare et al., 2009). Where the word is used in a subordinate sense for which only some of these features are relevant, strong initial activation of contextually irrelevant features will be followed by their continued cross-activation by relevant features. These two factors will jointly prevent complete suppression. Contextually irrelevant, but unsuppressed features will continue to support inferences. These inferences, supported by the dominant sense, but irrelevant for the subordinate use, engender fallacies of equivocation. This logic motivates the *linguistic salience bias hypothesis* (Fischer & Sytsma, 2021):

H0 Where unbalanced irregular polysemes have dominant uses that over-specify the information relevant for interpreting a subordinate use, this subordinate use will lead to fallacies of equivocation: It will trigger inappropriate inferences supported only by the dominant sense *and* these will influence further cognition, regardless of context.

E.g., Fischer and Engelhardt (2020) provided evidence from pupillometry and plausibility ratings that spatial inferences from the verb 'S sees X' to *X is in front of S*, that are supported only by the dominant visual sense, are also made from purely epistemic uses ('Joe saw the risks').

Extant studies. Since irregular polysemes need not form a homogenous class (Carston, 2021), it seems prudent to examine **H0** not 'at one go', with studies considering a wide variety of such words, but 'step by step', for restricted classes of similar words. This approach is common in experimental philosophy (X-Phi): Philosophers are typically interested in the semantic and processing properties of specific words of philosophical interest. Thus, most studies in X-Phi focus on specific words of interest (Sytsma & Livengood, 2015).

X-Phi provides the research context for extant work on **H0**. Philosophical discourse systematically employs familiar polysemes from ordinary discourse in subordinate (technical) uses. Philosopher J.L. Austin (1962) suggested that this practice gives rise to fallacies of equivocation in influential philosophical arguments including the arguments 'from illusion' and 'from hallucination'. This suggestion guided extant studies assessing **H0**. These studies implemented the psycho-linguistic cancellation paradigm with plausibility ranking and rating tasks, combined with reading-time measurements (Clifton et al., 2016) or pupillometry (Sirois & Brisson, 2014), and considered inferences from expressions employed in those arguments, viz, appearance- and perception-verbs (Fischer & Engelhardt, 2016; 2017; 2020; Fischer et al., 2021; Fischer, Engelhardt & Sytsma 2021). They provided evidence of inappropriate inferences that influence subsequent cognition – even from professional philosophers (Fischer, Engelhardt, & Herbelot, 2022).

These studies have one crucial limitation: In their materials, the polysemous word *precedes* the disambiguating context that supports its subordinate interpretation. It is therefore no surprise that the word initially triggers the inference of interest; **H0** is merely supported by the finding that this inference subsequently influences further judgment, despite its defeat by post-verbal context. This limitation has us ask:

RQ1. Will the subordinate use also trigger the inference of interest where the disambiguating context comes first, and invites the subordinate interpretation *from the start*?

RQ2. And, if so, will this inference still be strong enough to influence subsequent judgment and reasoning?

Appearance verbs. To address these questions in line with the X-Phi approach, we consider inferences from appearance verbs. Philosophical analysis (Brogaart, 2013; 2014) suggests that, in their dominant sense, appearance verbs are used to attribute belief, knowledge, and experience to their patients ('Jack looks dirty to Jane' ≈ Jane believes, indeed knows, and experiences, that Jack is dirty). Distributional semantic analysis of a parsed Wikipedia snapshot (Flickinger et al., 2010) revealed that the distributionally most similar words to 'appear', 'seem', and 'look' are doxastic verbs ('think', 'believe', 'find(mental)'), followed by epistemic verbs, while experiential verbs lacked prominence (Fischer, Engelhardt, & Herbelot, 2015). This suggests that, in their dominant use, appearance verbs are associated with a complex situation schema into which doxastic, epistemic, and experiential patient features are integrated with decreasing strength. Philosophers, however, often employ appearance verbs in a subordinate 'phenomenal' sense, to characterize the patients' subjective experience, without implications about their beliefs or knowledge (Chisholm, 1957; Robinson, 1994). This use is familiar to ordinary speakers (Fischer, Engelhardt, & Sytsma 2021, App.3). It can be interpreted by retaining the experiential patient features from the complex situation schema associated with the dominant use, while suppressing precisely the most strongly integrated features. In response to **RQ1-2**, **H0** thus motivates the verb-specific hypothesis.

H1 Subordinate phenomenal uses of appearance verbs ('look', 'appear', 'seem') trigger contextually inappropriate belief inferences supported only by the verbs' dominant sense, and these inferences influence further cognition – even when the verbs are *preceded* by a disambiguating context that invites phenomenal interpretation *from the start*.

Three studies

Approach and predictions. To assess **H1**, we conducted three experiments with the cancellation paradigm. In all three, participants read and rated the plausibility of three-sentence items like the following:

(1) The vessels waited far out at sea[1]. They looked[2] small[3] to Eve[4]. She thought they were big[5].
[1]Pre-verbal context [2]Source verb [3]Source adj. [4]Source object [5]Conflict adjective

According to **H1**, the appearance verb in S2 triggers a default belief inference from 'X looks F to S' to *S believes X is F* (e.g., in (1), *Eve believed the vessels were small*). The post-verbal context, in S3, is manipulated to be either consistent with this inference, or inconsistent (as in (1)). **H1** predicts:

(i) higher re-reading times for the 'source region' of the inference ('looks small to Eve') and the 'conflict region' ('were big'), in the inconsistent than the consistent condition (INCON > CON in re-reading times).

(ii) lower plausibility ratings, in the inconsistent than the consistent condition (INCON < CON).

(i) provides evidence that the inference is triggered, (ii) that the inference persists to influence subsequent judgments.

Crucially, **H1** predicts these consistency effects also for items with pre-verbal contexts that invite phenomenal interpretations from the start. To test this prediction, S1 provides a pre-verbal context that invites either dominant or phenomenal interpretations, or neither, by specifying different viewing conditions. Familiar conditions of veridical perception (where everybody believes/knows that objects look the size, shape, or color that they actually are) invite dominant interpretations (which include the default belief attribution to the patient). Familiar conditions of non-veridical perception (where everybody believes/knows that objects look a different size, shape, or color than they in fact are) invite phenomenal interpretations. Neutral conditions speak neither for nor against belief attributions and thus invite neither dominant nor phenomenal interpretations. Note that, in familiar non-veridical contexts, the default belief attribution is not only *unsupported by the intended phenomenal sense* of 'look', etc. It is also *wrong*, namely, *epistemically deviant*: It is, e.g., objectively improbable that a viewer will believe the vessels are small, just because they look small from this great distance; therefore, the belief attribution has a high probability of being false, in non-veridical contexts.

The preliminary Exp.1 assessed Prediction (ii) only, with a plausibility rating task. Exp.2-3 assessed both predictions, by combining such ratings with fixation time measurements. To be able to pick up potential contrast effects on processing and ratings of the crucial non-veridical items, each experiment paired this key condition with only one other veridicality condition. Exp.1 and Exp.3 paired non-veridical items with neutral items. Exp.2 paired non-veridical with veridical items. Exp. 3 was pre-registered on OSF.

Experiment 1

Exp.1 Method

Participants. 175 UK participants (57% female, 1 non-binary; 93% of participants aged 15-35), screened for first language English and possession of a degree, were recruited via *Prolific* to complete a *Qualtrics* questionnaire.

Materials and procedure. A **norming study** served to assign critical three-sentence items (like (1) above) to veridicality-conditions. 200 participants (demographics matching main study) read 36 draft scenarios (S1-S2) using the verb 'look', which may have the weakest belief implications (Fischer et al. 2021). All scenarios involved visual objects and visual properties (size, shape, color). E.g.:

The vessels waited far out at sea. They looked small to Eve.

Participants then rated on a 7-point scale ('-3' to '+3'), how confident they were that the object viewed actually had the property it appeared, e.g.:

The vessels were small.

We considered whether participants were confident (mean ratings significantly above neutral mid-point) that the statement was true (so that the scenario specified familiar veridical viewing conditions) or confident that the statement was false (so that the scenario specified familiar non-veridical viewing conditions, as in the 'vessel' example), or neither (mean ratings not significantly different from mid-point). We assigned items to veridicality conditions (veridical, non-veridical, neutral), accordingly.

In the main study, participants read 69 items, including 36 critical items which had either non-veridical or neutral pre-verbal contexts, and rated their plausibility on a 7-point scale (from 'very implausible' (-3) to 'very plausible' (+3)).

Design and analysis. We used a 2×3×2 within-subjects design, manipulated veridicality (non-veridical vs neutral, in S1), verb ('look', 'appear', 'seem', in S2), and consistency with the belief inference (CON vs INCON, in S3), and analyzed data with a repeated-measures ANOVA.

Exp.1 Results

We found a significant veridicality by consistency interaction ($F(1,174) = 79.62$, $p < .001$; $\eta^2 = .314$) (see Fig.1), a medium-sized main effect of veridicality ($F(1,174) = 17.69$, $p < .001$; $\eta^2 = .092$), and a large effect of consistency ($F(1,174) = 64.39$, $p < .001$; $\eta^2 = .27$).

As predicted, mean ratings for consistent items were higher than for inconsistent items, in both the neutral condition ($t(174) = 10.57$, $p < .001$, Cohen's $d = .80$) and the non-veridical ('negative' in Fig.1) condition ($t(174) = 4.573$, $p < .001$, $d = .35$). This small effect is consistent with **H1**(ii) but motivates more rigorous examination of *both* our predictions.

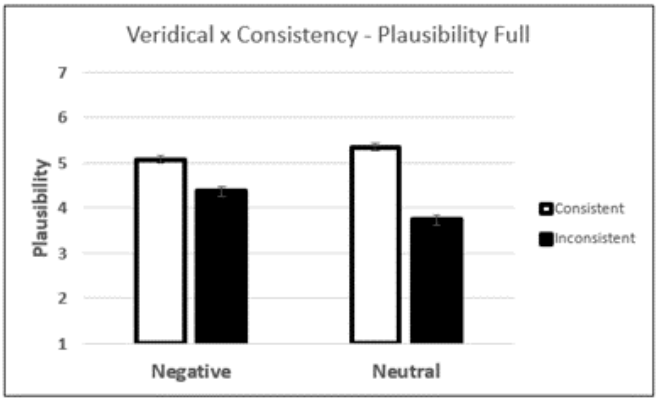

Figure 1. **Exp.1. Mean plausibility ratings** on 7-point scale.

Experiment 2

To obtain more and improved materials for Exp.2-3, we conducted a new **two-round norming study** via *Prolific* with *Qualtrics* (N1=100, N2=202) from a demographic matching the main study sample. Participants read 'look' versions of 61 scenarios (S1-2), again involving visual objects and the properties of size, shape, and color, and were given the same task as in the previous norming study. Scenarios were assigned to veridicality-conditions as before.

Prior further norming work ensured words had similar length across conditions in the regions of interest. We also used frequency information for British English (Leech, Rayson, and Wilson, 2001), to ensure the mean frequencies of 'source' and 'conflict adjectives' in relevant conditions were similar. Table 1 presents sample normed items that were used in Exp.2-3 for each veridicality condition.

Table 1. Sample items.

Non-veridical ('negative')	The fishing rod was immersed in the water. The rod looked bent to the fisherman. He thought it was bent/straight.
Veridical ('positive')	The visitor stood in front of the house entrance. He seemed tall to the host. She believed he was tall/short.
Neutral	The lighting in the room was odd. The hostess's dress looked blue to Hannah. She thought it was blue/green.

Exp.2 Methods

Participants. 45 undergraduate psychology students (1st and 2nd year) from the University of East Anglia participated for course credit. All were native speakers of British English with normal or corrected-to-normal vision.

Materials and procedure. Participants read and rated 96 items, including 48 critical items, presented in a randomized order. Eye movements (from the right eye) were recorded with an SR Research Ltd. EyeLink 1000 eye-tracker and head movements were minimized with a chin rest. We measured first-pass and re-reading times for five regions of interest, indicated in (1) above. After a 9-point calibration and validation procedure, participants completed 4 practice trials and 96 experimental trials. Each participant saw an equal number of items in each condition, as verbs were rotated across items using a Latin Square Design. Before each trial, participants fixated a drift-correction dot on the edge of the monitor, centered vertically. The item appeared after an interval of 500 ms, its initial letter always displayed in the same position as the drift correction dot. The entire item appeared on a single line on the screen, presented in 12 pt. Participants read each item silently and then pressed the spacebar on the keyboard. A plausibility-rating prompt appeared; participants rated items' plausibility on a 5-point scale, by pressing the corresponding key on the keyboard.

Design and analysis. In a 2×3×2 within-subjects design, we manipulated veridicality (veridical or 'positive' vs non-veridical or 'negative', in S1), verb ('look', 'appear', 'seem', in S2) and consistency with the belief inference (CON vs INCON, in S3). To analyze results, we ran linear mixed effects (LME) models in R (Bates et al. 2018; R Core Team 2018), including context, verb, and consistency as fixed effects and subjects and items as random effects.

Exp.2 Results

Findings from Study 2 fully confirmed predictions for plausibility ratings (INCON < CON, *even for non-veridical items,* if in attenuated form) (see Fig. 2) and for rereading times (INCON > CON, in source and conflict regions).

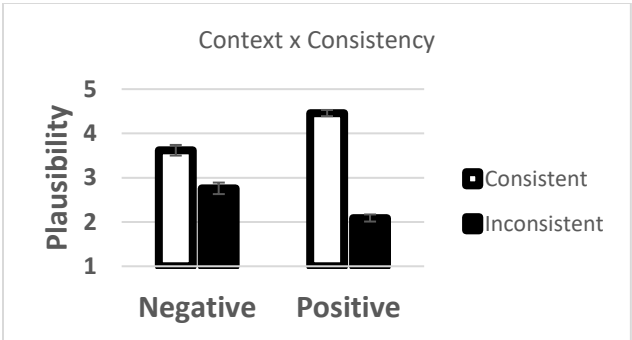

Figure 2. **Exp.2. Mean plausibility ratings** on 5-point scale (1 = 'very implausible', 5 = 'very plausible'). Error bars show the standard error of the mean.

Plausibility ratings. We found a context by consistency interaction ($t = 7.42$, $p < .001$) and a main effect of consistency ($t = 15.85$, $p < .001$). All paired comparisons were significant ($p < .001$). One sample t-tests (computed on participant means) showed that ratings in consistent conditions were significantly above mid-point (positive: $t(45)$

= 20.84, $p < .001$, $d = 3.07$; negative: $t(45) = 5.12$, $p < .001$, $d = .76$), while significantly or marginally below mid-point, in the inconsistent conditions (positive: $t(45) = -11.80$, $p < .001$, $d = -1.74$; negative: $t(45) = -1.83$, $p = .074$ ($d = -.27$).

Re-reading times. We found the predicted consistency effects in three regions of interest (source verb: $t = -2.65$, $p = .011$; source object: $t = -2.15$, $p = .037$; conflict adjective: $t = -2.43$, $p = .019$), and a marginal effect in a fourth (source adjective: $p = .089$). For these regions, **H1** predicted higher rereading times INCON > CON conditions, which we found.

Experiment 3

Exp.3 Methods

Participants. 48 psychology undergraduates (1st and 2nd year) from the same institution, meeting the same restrictions, participated for course credit.

Materials and procedure were the same as in Exp.2, except that, in the critical items, the same non-veridical items were now used alongside neutral (instead of veridical) items.

Design and analysis were the same as in Exp.2, except for the veridicality manipulation (now non-veridical vs neutral).

Exp.3 Results and discussion

Plausibility ratings. Unlike Exp.1-2, Exp.3 did not find a difference in mean plausibility ratings between consistent and inconsistent items in the key non-veridical ('negative') condition, while the difference materialized again in the neutral condition (see Fig.3).

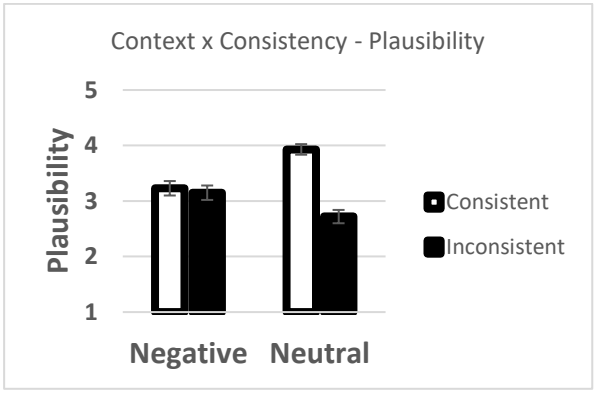

Figure 3. **Exp.3 Mean plausibility ratings** on 5-point scale. **Whole sample** (N=48). Error bars show the standard error of mean.

We asked whether these whole-sample means might mask different response patterns between 'correct' responders (correctly judging CON ≤ INCON items, in the non-veridical conditions) and 'biased' responders (who incorrectly deem CON > INCON, even in this condition). The according assignment of participants to two groups revealed an almost even split between 26 'correct' and 22 'biased' responders (see Fig. 4 below). To account for this potential group difference, we included an additional 'group' variable in the LME model. Results showed a significant main effect of consistency $t = -7.27$, $p < .05$ (CON > INCON), an interaction between context and consistency $t = 11.65$, $p < .05$, and a 3-way interaction between context, consistency, and group $t = -4.76$, $p < .05$. We followed up this 3-way interaction with paired comparisons for each group. All but one were significant (p's <.05, most p's <.001); however, the correct responders deemed consistent and inconsistent items equally plausible, in the neutral condition ($p > .5$).

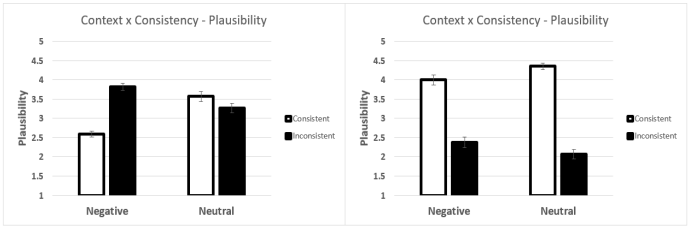

Fig. 4. **Exp.3 Mean plausibility ratings by group** on 5-point scale. 26 correct responders left, 22 biased responders right.

Re-reading times. To examine whether these group differences are genuine or due to noise, we added a variable for group also to the LME analyses of reading times. We found the consistency effects (INCON > CON) predicted by H1 in rereading times in 3 regions of interest (source adjective: $t = -4.35$, $p < .05$, source object: $t = -2.43$, $p < .05$, conflict adjective: $t = -4.01$, $p < .05$). We also found a main effect of group precisely in the source region from which the inappropriate inference originates (source verb: $t = -2.16$, $p < .05$; source adjective: $t = -2.24$, $p < .05$.): The 26 'correct responders' reread these regions more extensively than the 22 'biased responders' (source verb: N=26: 263ms vs. N=22: 202ms; source adjective: N=26: 240ms vs. N=22: 182ms). We inferred that the correct responders initially make the default belief inferences, just like the biased responders, but then make more effort to suppress it, leading to more successful suppression, which aligns plausibility ratings with the level of contextual support for the belief inferences. Rereading times thus suggest group differences in plausibility ratings are genuine, rather than due to noise.

Discussion. The two different response patterns had us look for similar group differences in Exp.2. There, however, only 4 of 45 participants (8.9%) displayed the correct response pattern in the non-veridical condition. The only difference between the two experiments is that the same non-veridical items are contrasted with veridical items in Exp.2 and with neutral items in Exp.3. We inferred that judgments about the key non-veridical items are influenced by what items they are contrasted with. We therefore reanalyzed data from Exp.1, which (like Exp.3) employed non-veridical and neutral items. We found the whole-sample means had masked the exact same two response patterns (see Figure 5).

As in Exp.3, all but one paired comparisons were significant (p's ≤ .007), with the same exception: Correct responders again deemed consistent and inconsistent items

equally plausible, in the neutral condition ($p = .86$). In the key non-veridical condition, consistency effects were again large for both correct responders $t(63) = -10.146$, $p < .001$, Cohen's $d = -1.268$ and biased responders $t(110) = 14.574$, $p < .001$, $d = 1.383$, with the correct responders 'flipping' the biased response pattern. These patterns were displayed by different proportions of participants, in the two experiments: Biased responders were in the clear majority (63.4%) in Exp.1, but just shy of half (45.8%) in Exp.3.

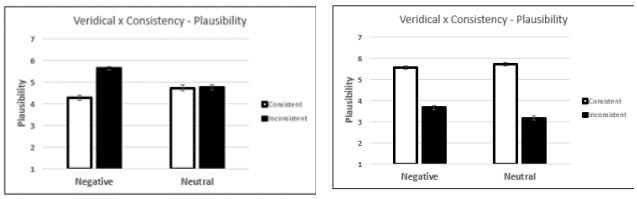

Fig. 5. **Exp.1 Mean plausibility ratings by group**, on 7-point scale. 64 correct responders left, 111 biased responders right.

The neutral items are the most difficult to judge, since their context does not support any assessment of the belief attributions, either way. Such difficulty ('answer disfluency') prompts analytic processing (Alter et al., 2007; 2013; Thompson et al., 2013). We tentatively infer that these items functioned as reflection prompts, supporting critical scrutiny and suppression of contextually inappropriate automatic inferences, across the board. These prompts may be yet more effective in a lab setting (as in Exp.3) than in the everyday settings in which *Prolific* studies (like Exp.1) are taken. More reflective participants then make more effort to suppress contextually unsupported and epistemically deviant default inferences, and are more successful. This interpretation predicts that overall reading times are highest for the (most difficult and disfluency-inducing) neutral items and that these dwell times are higher for non-veridical items when they are paired with ('reflection prompting') neutral items than when paired with veridical items. A follow-up analysis confirmed this was the case (Table 2).

Table 2. **Mean dwell times in Exp.2-3**

Veridicality condition	*Ms*
Neutral	857
Non-veridical (vs. neutral)	755
Non-veridical (vs. veridical)	710
Veridical	706

General Discussion

Main findings. We found qualified support for the linguistic salience bias hypothesis (**H0**) that specifies a set of processing conditions under which default inferences beat contextual information: This happens when markedly unbalanced polysemes have a dominant sense that is associated with a schema that includes but over-specifies the semantic information relevant for interpreting subordinate uses and these are interpreted with the Retention/Suppression Strategy. We studied inferences from appearance verb, which we argued meet these conditions, and found:

(1) Re RQ1: Evidence from re-reading times suggests that subordinate (phenomenal) uses of appearance verbs trigger default (belief) inferences that are supported only by their dominant sense, even when pre-verbal contexts (specifying non-veridical viewing conditions) invite subordinate (phenomenal) interpretation from the start.

(2) Re RQ2: Evidence from plausibility ratings suggests that these contextually inappropriate and epistemically deviant automatic inferences influence further cognition, though this influence is mitigated by reflection prompts (viz., difficult contrast items that engender 'answer disfluency').

(3) Upon inclusion of reflection prompts, group differences emerged in both re-reading times and plausibility ratings: When prompted to reflect, up to roughly half of participants managed to disregard the default inferences, in their further judgment and reasoning, and to successfully align their judgments with the levels of contextual support available for the default inference. Even then, however, the default inferences of interest strongly influenced subsequent judgments of roughly half of participants.

Limitations and future directions. These studies on appearance verbs contribute to an incremental approach that examines the linguistic salience bias hypothesis for one restricted word class after the other. They motivate use of the same experimental paradigm to examine inferences from other unbalanced polysemes.

Whereas findings (1)-(2) were made by examining initial hypotheses, we arrived at (3) through interpretation of results, supported by post-hoc analyses. Accordingly, (3) should be treated as a hypothesis to be examined by future work, using well-understood analytic thinking primes and reflection prompts (cf. Hertwig & Ortmann, 2001; Lerner & Tetlock, 1999). Group differences emerging upon inclusion of reflection prompts may be due to individual differences in reflectiveness (Frederick, 2005) or the ability to suppress pre-potent responses (inhibition; Hasher et al., 2007). Individual differences studies can thus help assess present conclusions.

Philosophical and wider relevance. The linguistic salience bias posited by **H0** poses a challenge to philosophical practice: Philosophers frequently give subordinate (regimented, technical) uses to familiar terms. The emerging field of 'conceptual engineering' explicitly examines how familiar words can be optimized for research purposes or to change attitudes, in pursuit of societal agendas (Cappelen & Plunkett, 2020). The field has begun to consider empirical constraints on linguistic innovation (Fischer, 2020; Machery, 2021). The linguistic salience bias will lead to fallacies in reasoning with new or '(re-)engineered' subordinate senses, where the polyseme at issue is strongly unbalanced and subordinate uses are interpreted through Retention/ Suppression. Our studies support this challenge, but simultaneously suggest it may be mitigated by reflectiveness.

References

Alter, A.L., Oppenheimer, D.M., Epley, N., & Eyre, R.N. (2007). Overcoming intuition: Metacognitive difficulty activates analytic reasoning. *Journal of Experimental Psychology: General, 136*(4), 569-576.

Alter, A.L., Oppenheimer, D.M., & Epley, N. (2013). Disfluency prompts analytic thinking- But not always greater accuracy: Response to Thompson et al. (2013), *Cognition*, *128*(2), 252-255.

Austin J.L. (1962). *Sense and Sensibilia*. Oxford: Oxford University Press.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1-48.

Bicknell, K., Elman, J.L., Hare, M., McRae, K., & Kutas, M. (2010). Effects of event knowledge in processing verbal arguments. *Journal of Memory and Language, 63,* 489-505.

Brocher, A., Foraker, S. & Koenig, J.P. (2016). Processing of irregular polysemes in sentence reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition 42*: 1798-1813. https://doi.org/10.1037/xlm0000271.

Brocher, A., Koenig, J.-P., Mauner, G. & Foraker, S. (2018). About sharing and commitment: the retrieval of biased and balanced irregular polysemes. *Language, Cognition and Neuroscience*, *33*: 443-466. https://doi.org/10.1080/23273798.2017.1381748.

Brogaard, B. (2013): It's not what it seems. A semantic account of 'seems' and seemings, *Inquiry 56*, 210-239.

Brogaard, B. (2014). The phenomenal use of 'look' and perceptual representation. *Philosophy Compass 9*: 455-468. https://doi.org/10.1111/phc3.12136.

Byrd, R.J., Calzolari, N., Chodorow, M.S., Klavans, J.L., Neff, M.S. & Rizk, O.A. (1987). Tools and methods for computational lexicology. *Computational Linguistics 13*: 219-40. https://dl.acm.org/doi/10.5555/48160.48163.

Cappelen, H. & Plunkett, D. (Eds.) (2020). *Conceptual Engineering and Conceptual Ethics*. Oxford University Press.

Carston, R. (2021). Polysemy: Pragmatics and sense conventions. *Mind & Language*, *36*(1), 108-133.

Chisholm, R.M. (1957). *Perceiving: A philosophical study.* Ithaca: Cornell University Press.

Clifton, C., Ferreira, F., Henderson, J.M., Inhoff, A.W., Liversedge, S.P., Reichle, E.D. & Schotter, E.R. (2016). Eye movements in reading and information processing: Keith Rayner's 40 year legacy. *Journal of Memory and Language 86*: 1-19. https://doi.org/10.1016/j.jml.2015.07.004.

Elman J.L. & McRae, K. (2019). A model of event knowledge. *Psychol Rev. 126*(2): 252-291. doi: 10.1037/rev0000133. Epub 2019 Jan 31. PMID: 30702315.

Faust, M.E., &Gernsbacher, M.A. (1996). Cerebral mechanisms for suppression of inappropriate information during sentence comprehension. *Brain and Language 53*: 234-259. https://doi.org/10.1006/brln.1996.0046.

Ferretti, T.R., McRae, K., & Hatherell, A. (2001). Integrating verbs, situation schemas, and thematic role concepts. *Journal of Memory and Language, 44*, 516-547.

Ferretti, T.R., Kutas, M., & McRae, K. (2007). Verb aspect and the activation of event knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33,* 182-196.

Fischer, E. (2020). Conceptual control. On the feasibility of conceptual engineering. *Inquiry*. 2020, 1-29. http://dx.doi.org/10.1080/0020174X.2020.1773309.

Fischer, E. & Engelhardt, P.E. (2016). Intuitions' linguistic sources: Stereotypes, intuitions, and illusions. *Mind and Language 31*: 67-103. https://doi.org/10.1111/mila.12095.

Fischer, E. & P.E. Engelhardt. 2017. Stereotypical inferences: Philosophical relevance and psycholinguistic toolkit. *Ratio* 30: 411-442. https://doi.org/10.1111/rati.12174.

Fischer, E. & Engelhardt, P.E. (2020). Lingering stereotypes: Salience bias in philosophical argument. *Mind and Language 35*: 415-439. https://doi.org/10.1111/mila.12249.

Fischer, E., Engelhardt, P.E. & Herbelot, A. (2015). Intuitions and illusions: From experiment and explanation to assessment. In *Experimental Philosophy, Rationalism and Naturalism*, ed. E. Fischer & J. Collins, 259-292. London: Routledge.

Fischer, E., Engelhardt, P.E., & Herbelot, A. (2022). Philosophers' linguistic expertise: A psycholinguistic approach to the expertise objection against experimental philosophy. *Synthese*, *200*, 1-33. https://doi.org/10.1007/s11229-022-03487-3

Fischer, E., Engelhardt, P.E., Horvath, J. & Ohtani, H. (2021). Experimental ordinary language philosophy: A cross-linguistic study of defeasible default inferences. *Synthese 198*: 1029-1070. https://doi.org/10.1007/s11229-019-02081-4.

Fischer, E., Engelhardt, P.E. & Sytsma, J. (2021). Inappropriate stereotypical inferences? An adversarial collaboration in experimental ordinary language philosophy. *Synthese 198*: 10127-10168. https://doi.org/10.1007/s11229-020-02708-x.

Fischer, E., & Sytsma, J. (2021). Zombie intuitions. *Cognition 215*: e104807. https://doi.org/10.1016/j.cognition.2021.104807.

Flickinger, D., Oepen, S. & Ytrestol, G. (2010). Wikiwoods: syntacto-semantic annotation for the English Wikipedia. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner and D. Tapias (eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC2010)* (pp. 1665-1671). Paris: European Language Resources Association.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives* 19(4): 25-42.

Giora, R. (2003). *On Our Mind. Salience, Context, and Figurative Language.* Oxford: Oxford University Press.

Giora, R. (2012). The psychology of utterance processing: Context vs Salience. In: K. Jaszczolt & K. Allan (Eds.), *The Cambridge Handbook of Pragmatics* (151-167). Cambridge: Cambridge University Press.

Hampton, J. (2006). Concepts as prototypes. In *The psychology of learning and motivation: Advances in research and theory*, ed. Ross, B.H., 79-113. Amsterdam: Elsevier.

Hare, M., Jones, M., Thomson, C., Kelly, S. & McRae, K. (2009). Activating event knowledge. *Cognition 111* : 151-167. https://doi.org/10.1016/j.cognition.2009.01.009.

Hasher, L., Lustig, C., & Zacks, R. (2007). Inhibitory mechanisms and the control of attention. In A.R.A. Conway et al. (Eds.), *Variation in working memory* (pp. 227-249). Oxford University Press.

Hertwig, R., & Ortmann, A. (2001). Experimental practices in economics: A methodological challenge for psychologists? *Behavioral and Brain Sciences, 24*(03), 383-403.

Kamide, Y., Altmann, G. T. M., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, *49,* 133-156.

Kim, A. E., Oines, L. D., & Sikos, L. (2016). Prediction during sentence comprehension is more than a sum of lexical associations: the role of event knowledge. *Language, Cognition, and Neuroscience*, *31*, 597-601.

Klepousniotou, E., Titone, D. & Romero, C. (2008). Making sense of word senses: The comprehension of polysemy depends on sense overlap. *Journal of Experimental Psychology: Learning, Memory, and Cognition 34*: 1534-1543.

Klepousniotou, E., Pike, B. Steinhauer, K. & Gracco, V. (2012). Not all ambiguous words are created equal: an EEG investigation of homonymy and polysemy. *Brain and Language 123*: 11-21. https://doi.org/10.1016/j.bandl.2012.06.007.

Leech, G., Rayson, P. & Wilson, A. (2001). *Word Frequencies in Written and Spoken English: Based on the British National Corpus*. London: Routledge.

Lerner, J.S., & Tetlock, P.E. (1999). Accounting for the effects of accountability. *Psychological Bulletin, 125*(2), 255.

MacGregor, L.J., Bouwsema, J. & Klepousniotou, E. (2015). Sustained meaning activation for polysemous but not homonymous words: Evidence from EEG. *Neuropsychologia 68*: 126-138. https://doi.org/10.1016/j.neuropsychologia.2015.01.008.

Machery, E. (2021). A new challenge to conceptual engineering. *Inquiry*. https://doi.org/10.1080/0020174X.2021.1967190

Matsuki, K., Chow, T., Hare, M., Elman, J. L., Scheepers, C., & McRae, K. (2011). Event-based plausibility immediately influences on-line language comprehension. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *37,* 913-934.

Pylkkänen, L., Llinás, R. & Murphy, G.L. (2006). The representation of polysemy: MEG evidence. *Journal of Cognitive Neuroscience 18*: 97-109. https://doi.org/10.1162/089892906775250003.

R Core Team (2018). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria.

Robinson, H. (1994). *Perception*. London: Routledge.

Rosch, E. (1975). Cognitive representation of semantic categories. *Journal of Experimental Psychology: General 104*, 192-233.

Rumelhart, D.E. (1978). Schemata: The building blocks of cognition. In *Theoretical Issues in Reading Comprehension*, ed. R. Spiro, B. Bruce, and W. Brewer, Hillsdale, 33-58. NJ: Lawrence Erlbaum.

Sirois, S., & Brisson, J. (2014). Pupillometry. *WIREs Cognitive Science*, *5*, 679-692. https://doi.org/10.1002/wcs.1323.

Sytsma, J. & Livengood, J. (2015). *The Theory and Practice of Experimental Philosophy*. Peterborough, Ontario: Broadview Press.

Thompson, V., Prowse Turner, J., Pennycook, G., Ball, L. J., Brack, H., Ophir, Y., & Ackerman, R. (2013). The role of answer fluency and perceptual fluency as metacognitive cues for initiating analytic thinking. *Cognition*, *128*(2), 237-251.

Yee, E., Jones, M.N., & McRae, K. (2017). Semantic Memory. In J. T. Wixted & S. Thompson-Schill (Eds.), *The Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience* (4th Ed., Vol.3). New York: Wiley.

Zwaan, R.A. (2016). Situation models, mental simulations, and abstract concepts in discourse comprehension. *Psychonomic Bulletin & Review 23*: 1028-1034. https://doi.org/10.3758/s13423-015-0864-x.