Preprint version of paper forthcoming in: E. Fischer and M. Curtis (eds.) (2019). *Methodological Advances in Experimental Philosophy*. London: Bloomsbury.

Eyes as windows to minds: Psycholinguistics for experimental philosophy

Eugen Fischer and Paul E. Engelhardt

Abstract: Psycholinguistic methods hold great promise for experimental philosophy. Many philosophical thought experiments and arguments proceed from verbal descriptions of possible cases. Many relevant intuitions and conclusions are driven by spontaneous inferences about what else must also be true in the cases described. Such inferences are continually made in language comprehension and production. This chapter explains how methods from psycholinguistics can be employed to study such routine automatic inferences, with a view to assessing intuitions and reconstructing arguments. We demonstrate how plausibility ratings, pupillometry, and reading time measurements can be used to examine hypotheses about automatic inferences in speech and text comprehension. Two experiments on inferences from polysemous (perception-)verbs provide evidence of a potentially consequential 'salience bias'. Findings help assess intuitions about unusual cases and analyse a philosophical paradox ('argument from hallucination'). The paper thus illustrates how we can adapt psycholinguistic methods for philosophical purposes and demonstrates the methods' philosophical usefulness.

Keywords: eye-tracking, pupillometry, plausibility-ratings, automatic comprehension inferences, argument analysis, paradoxes about perception.

Much philosophical thought occurs in natural language, as thinkers read or write philosophical texts, discuss philosophical problems with each other, or engage in the subvocalized speech characteristic of conscious thought (Carruthers 2002). Philosophical thought is therefore bound to be influenced by the automatic processes that continually go on in language comprehension and production. Much philosophical reasoning proceeds from verbal descriptions of possible cases. In thought experiments, such descriptions prompt intuitions about what else is also true of the cases described, and such intuitive judgments are frequently treated as evidence for or against philosophical theories (review: Weinberg 2016, *pace* Cappelen 2012 and Deutsch 2015). Many philosophical arguments involve inferences from premises that describe a possible case, to conclusions about what else must also be true of it. Such judgments and conclusions can be generated by routine comprehension inferences, which, for example, have us automatically infer from 'The secretary fell out of the window' that the protagonist is female (Atlas and Levinson 1981), was initially located in a building, and was subsequently injured or killed (McKoon and Ratcliff 1980). While many inferences triggered by philosophical case descriptions may be due to domain-specific processes (like 'mindreading', which may generate intuitive knowledge attributions; see Nagel 2012, Gerken 2017), many others will be due to routine comprehension processes.

An important strand of experimental philosophy examines whether and when case intuitions have evidentiary value – and philosophers possess warrant for accepting them. Most of this work to date has proceeded by examining the sensitivity of relevant intuitions to truth-irrelevant parameters (like the order in which cases are presented) and infers lack of evidentiary value where it observes such sensitivity (reviews: Mallon 2016, Nichols and Knobe 2017, Stich and Tobia 2016). Partially in response to replication issues and theoretical challenges to the key inference from observed sensitivity to lack of evidentiary value (e.g. Horne and Livengood 2017), recent calls for an 'experimental philosophy 2.0' (Nado 2016) have suggested that research in this strand should (a) be refocussed on the examination of specific cognitive processes that underpin philosophical thought (building on, e.g. Nichols and Knobe 2007), (b) deploy the resulting understanding of how specific processes work, to develop 'epistemological profiles', which indicate under which conditions we may (not) trust the processes' outputs (Weinberg 2015, 2016), and (c) employ such profiles to assess a wider range of outputs: not only intuitive judgments but also inferences in arguments (Fischer and Engelhardt 2017a, 2017b).

This paper will discuss and demonstrate how experimental philosophers (and especially an 'experimental philosophy 2.0') can recruit methods from psycholinguistics to study automatic inferences in language comprehension and production, with a view to assessing philosophically relevant intuitions and arguments.[1] To demonstrate the approach and illustrate its potential philosophical usefulness, we will present a case study on the process of stereotypical enrichment (Levinson 2000) and its role in the influential 'argument from hallucination', a classical paradox about perception. This case study will contribute towards an epistemological profile of the key process: We will identify conditions under which stereotypical inferences predictably fail to lead to true conclusions, argue that these conditions obtain in formulations of the target argument, and present two studies – including one fresh study – which deployed different psycholinguistic methods to examine the hypothesis that under the conditions predicted, competent language users cannot help making inappropriate stereotypical inferences, despite knowing they are inappropriate (*sic*). The findings will support a novel resolution of the targeted philosophical paradox.

Section 1 will give an initial overview of the psycholinguistic methods that have been used to study automatic inferences involved in stereotypical enrichment and other comprehension/production processes. Section 2 will present our chosen philosophical application: It will identify a cognitive bias ('salience bias') besetting the generally reliable process of stereotypical enrichment, and explain how the processes' nascent epistemological profile can be used to assess philosophically relevant intuitions and arguments; a fresh analysis of the 'argument from hallucination' will suggest this classical paradox relies on stereotypical inferences which are contextually inappropriate. Section 3 will explain how we have used questionnaire-based methods and convenient 'offline' (outcome) measures to study comprehension inferences and garner first evidence of contextually inappropriate stereotypical inferences. The following sections will explore

[1] For use of psycholinguistic methods in conceptual analysis, see Powell et al. (2014).

how these methods can be complemented by 'online' measures (which tap into cognitive processes as they unfold). We will explore approaches that use people's eyes as windows into their minds: Section 4 will discuss a study that employs pupillometry (measurements of pupil dilations) to provide further evidence of inappropriate automatic inferences, in speech comprehension. Section 5 will report a fresh study that measures reading times to investigate inferences in text comprehension. Section 6 will present some potential methodological lessons.

1. Automatic inferences and their psycholinguistic study

We now introduce the automatic inference process of interest and then review the psycholinguistic methods that have been used to study it.

1.1. Stereotypical enrichment

Semantic memory is our memory for facts and 'general world knowledge', as opposed to personally experienced or 'episodic' events (McRae and Jones 2013, Tulving 2002). It is commonly conceived as a semantic network which doubles as information-storage and inference engine. In first approximation, such a network consists of nodes representing concepts and links between them that can automatically pass on activation from stimuli, verbal and other, along several pathways simultaneously (Allport 1985). When a concept is '*activated'* it is more likely to be used by several cognitive processes, crucially including processes involved in utterance comprehension (from word-recognition to disambiguation) and forward inference: Links in semantic memory facilitate a plethora of probabilistic parallel inferences, in processes including language comprehension.

According to standard conceptions of semantic memory (McRae and Jones 2013, Neely and Kahan 2001),[2] the observed co-occurrence of features (things and their common properties, wholes and their common parts) and events (causes and typical effects, etc.) forges links between the respective nodes which grow stronger upon frequent activation and atrophy upon disuse. The more frequently we encounter tomatoes that are red (in the supermarket) or Germans who are nasty (in war movies), the stronger the links between the respective concepts become, the more activation gets passed on from the stimuli 'tomato' and 'German', respectively, to nodes representing 'red' and 'nasty', respectively. These concepts thus come to be *stereotypically associated* with the words: They are activated most rapidly and strongly, and come to mind first, when we encounter the words. The strength of these associations encodes information about the co-occurrence frequencies in the subject's physical and discourse environment. Such empirical knowledge is brought to bear in processes including language comprehension: While stereotypical associations do not determine the extension of words (Hampton and Passanisi 2016), they support automatic inferences from words ('tomato', 'German') to stereotypically associated features (*red* and *nasty*, respectively).

Embedded in cooperative communication (Grice, 1989), such stereotypical inferences address the challenge of the 'communication bottleneck'. Articulation of

[2] Kahneman (2011, part I) provides an elegant introduction; textbook: Harley (2014, ch.11).

speech proceeds at a slower pace than pre-articulation processes in speech production (Wheeldon and Levelt 1995) or parsing- and inference-processes in comprehension (Mehler et al. 1993) – "inference is cheap, articulation is expensive" (Levinson 2000: 29). This bottleneck is mitigated by inferences which systematically draw on information encoded by stereotypes. This process of *stereotypical enrichment* is captured by the bi-partite *I-heuristic* ('What is expressed simply is stereotypically exemplified', Levinson 2000, 37): Speakers facilitate and listeners devise interpretations that are positive, stereotypical, and highly specific, in line with the maxims:

(I-speaker) Skip mention of stereotypical features but make deviations from stereotypes explicit (e.g. 'male secretary').

(I-hearer) In the absence of such explicit indications to the contrary, assume that the situation talked about conforms to the relevant stereotypes, deploy (also) the most specific stereotypes relevant, and fill in details in line with this knowledge about situations of the kind at issue.

We now review some of the methods psycholinguists have used to study stereotypical associations and the automatic inferences they support, alongside some of the key findings.

1.2. Psycholinguistic methods

Stereotypical associations have been examined through a variety of offline and online measures. Offline measures include plausibility-ratings (How plausible/common is it that a tomato is red?), how frequently and early features are mentioned in listing tasks (List common features of tomatoes!), and the frequency with which participants use a word to complete a sentence-frame ('Tomatoes are___') (*cloze probability*) (McRae et al. 1997).

Priming experiments then serve to examine activation: Participants are presented with a '*prime'* word, sentence (-fragment), or short text and then a '*probe'* word or letter string, and have to, e.g. read out the word, decide whether the string forms a word, or judge whether the referents of prime and probe words both fall under the same category (e.g. *good* or *bad*). That the prime activates the probe concept is inferred from shorter response times, e.g. for 'money'-'bank' than 'honey'-'bank' (Lucas, 2000). Varying the time between the presentation of prime and probe ('stimulus onset asynchrony') allows researchers to examine the time course of activation.

Priming experiments have shown that single words, presented in isolation, activate stereotypically associated features rapidly, within 250 milliseconds (review: Engelhardt and Ferreira 2016). Event-nouns (Hare et al. 2009) and verbs (Ferretti et al. 2001) activate a particularly wide and complex range of features: Where the actions or events denoted typically involve particular kinds of agents, 'patients' acted on, instruments used, or relations between them (Tanenhaus et al. 1989), event words activate typical features pertaining to the fillers of all these different thematic roles. For example, the verb 'dig' activates the instrument *spade* (Ferretti et al., 2001), 'arrest' activates the agent *cop* and the patient *criminal* (ibid.), while telic verbs ('washing') swiftly activate both initial and resulting patient properties (*dirty*, *clean*) (Welke et al. 2015). Conversely, typical agent-,

patient-, instrument-, and location-nouns activate relevant verbs (Hare et al. 2005). Different words thus activate comprehensive stereotypes that can include perceptual features (e.g. *small*, *dirty*), evaluative features (e.g. *mean*) and functional features (i.e. involvement of particular instruments) (McRae et al., 1997, Ferretti et al. 2001).

Patterns of activation of features by words thus represent comprehensive 'event knowledge' about the typical features of actions like sewing and washing: who typically does it, with what instruments, what consequences are typically caused or intended, etc. Stereotypes can represent such knowledge because they are not unstructured 'bags of features' but have internal (thematic) structure. In incremental language comprehension, activation of features is sensitive to thematic roles (agent, patient, etc.): Sentence fragments ('She was arrested by the ___') activate typical agents (*cop*) in post-verbal position only when they leave the agent role blank (as above), not when they leave open the patient ('She arrested the ___'), and *vice versa* (Ferretti et al. 2001: Exp.4). These complex, internally structured stereotypes are known as *generalized situation schemas* (Rumelhart 1978, Tanenhaus et al. 1989).

Various online measures have been used to examine the inferences from words potentially supported by schemas (e.g. from 'S is sewing' to *S uses a needle*). These studies typically use a *'cancellation paradigm'* and materials where the word of interest is followed by a sequel that is inconsistent with (cancels) the hypothesised inference (e.g. '… the job would be easier if Carol had a needle'; see Harmon-Vukic et al. 2009). If participants make the critical inference from the prior verb, their conclusion's clash with the sequel will create comprehension difficulties, which require effort to overcome. This effort is reflected in different measures.

Pupillometry exploits the fact that cognitive effort makes our eyes widen, so that pupil dilation is an index of cognitive effort (Kahneman 1973, Laeng et al. 2012). We can therefore examine whether participants make automatic inferences from words by measuring their pupil size during and after they hear sentences with sequels that are either consistent or inconsistent with (cancel) the hypothesised inference from previous text. Dilations prompted by 'inconsistent' sentences but absent from otherwise identical 'consistent' counterparts are evidence of hypothesised inferences.

Reading-time measurements can build on the fact that when we read (rather than hear) sentences, comprehension difficulties cause us to slow down, and trigger increased backward (right-to-left) eye movements, called 'regressions'. The simplest way to detect slow-downs is to present text in small instalments of single words, sentences, or lines, on a computer screen, and ask participants to read at a comfortable pace and advance the text by pressing a key on the keyboard. Studies using this '*self-paced moving window'* paradigm show that participants read the remainder of the sentence more slowly when subject and verb were followed by a patient atypical for that agent-action pairing, rather than a typical patient ('The *mechanic*/*journalist* checked the spelling of his latest report') (Bicknell et al. 2010, Exp.1). A similar finding was made for instruments ('Susan used the *saw*/*scissors* to cut the expensive paper…'), despite the absence of single-word priming of typical patients (e.g. *scissors-paper*) (Matsuki et al. 2011). These findings suggest that readers make automatic inferences supported not only by stereotypes

associated with individual words ('journalist', 'mechanic', 'checking', etc.), but also by more specific situation schemas encoding knowledge (e.g. about car inspections) which are not associated with any one word but get activated by combinations of words (see also Metusalem et al. 2012).

Eye-trackers record the position of the reader's eye (up to every millisecond). They permit more fine-grained and differentiated reading-time measurements than the self-paced moving window paradigm as well as analyses of regressions. Dependent measures employed in the study of automatic inferences include[3]

- *first-pass reading time:* the sum of all fixations in a region of text, from first entering that region until leaving that region either in a forward or backward direction;
- *regression path duration:* the time from first entering a region until moving past that region forward in text (unlike first-pass reading time this also includes time spent on regressions out of the region);
- *second-pass reading time:* the sum of all fixations in a region following the initial first-pass reading; and
- *total reading times:* the sum of all fixations in a region.

A classic study (Rayner et al. 2004) examined inferences to typical patients prompted by combinations of verbs and instrument-nouns (consistent with the sequel in 1, but not 2):

(1) John used a knife to chop the large carrots for dinner last night.
(2) John used an axe to chop the large carrots for dinner last night.

Rayner and colleagues observed (marginally) increased regression-path durations on the (one-word) region of interest ('carrots') and (significantly) for the following (n+1) region ('for dinner'), in (2) than in (1). By contrast, where the word of interest was inconsistent with inferences prompted by the prior verb alone ('inflate' in 3, below), they observed also longer first-pass reading times (gaze durations) for both that word and its sequel.

(3) John used a pump to inflate the large carrots for dinner last night.

Inferences with different support (here: schemas activated by the verb alone or by combinations of verbs and nouns) or leading to expectation violations of different magnitude (3 > 2) may thus affect different eye-tracking measures.

Electroencephalogram (EEG) measurements often complement approaches that rely on the responsiveness of our eyes to clashes between expectations and subsequent text: In addition to the characteristic eye responses just reviewed, there are signature

[3] Unfortunately, definitions of these dependent variables differ across research labs and software packages. The present definitions are from Clifton et al. (2007) and employed in the current study (see Section 5).

electrophysiological responses in the brain, known as *'event-related brain potentials' (ERPs)* (Kutas and Federmeiner 2001, 2011). EEG measurements record electrical brain activity at the human scalp. By averaging, researchers can extract from such recordings a time series of changes in electrical brain activity before, during, and after an event of interest. The amplitudes, latencies, and scalp topographies of these evoked potentials were found to systematically vary not only with features of the linguistic or other stimulus but also with readers'/hearers' expectations. For example, violations of expectations based on syntactic rules (say, failure in gender agreement between pronoun and antecedent) produce positive deflections in the ERP waveform that peak 600 ms after stimulus onset (known as 'P600'). By contrast, violations of expectations based on knowledge encoded by stereotypes and schemas produce negative deflections that peak 400 ms after stimulus onset (known as 'N400').

ERP findings help us interpret the results of priming and eye-tracking studies: Semantic category violations ('Dutch trains are *sour*…') and conflicts with empirical knowledge ('Dutch trains are *white*…'; when they are actually yellow) lead to the same N400 response (with similar amplitude, topography, onset, and peak latency) (Hagoort et al. 2004). This suggests that lexical and empirical knowledge is deployed in the same way at the level of associative processing – and may both be encoded together as components of the same schema. Second, ERP results provide subtle further evidence that schemas are not 'bags of features' but deployed in a way sensitive to thematic roles: While N400 amplitudes indicate that the verb is expected where preceded by subject and object ('The restaurant owner forgot which *customer* the *waitress* had <u>served</u>…'), even where their typical roles are reversed ('…which *waitress* the *customer* had <u>served</u>') (Chow et al. 2016), this reversal prompts signature responses to syntactic violations (P600) suggesting participants expected the verb in the passive voice ('which waitress the customer had been served by'), consistent with assignments of agent and patient-roles typical for the verb (Kim et al. 2016, *cf.* Kim and Osterhout 2005). Finally, ERP studies suggest that inferences prompted by combinations of verbs and preceding nouns are supported by more specific situation schemas that are activated already at the verb (Bicknell et al. 2010, Exp.2).

The research reviewed supports a *'cued schemas account'* of language comprehension and production: 'Linguistic coding is to be thought of less like definitive content and more like interpretative clue' (Levinson 2000, 29). Words and syntactic constructions (Goldberg 2003, verb aspect: Ferretti et al. 2007, Kehler et al. 2008) are used as complementary clues for indicating and accessing relevant semantic and empirical knowledge in incremental language comprehension and production (Elman 2009). Relevant knowledge is encoded by situation schemas and other stereotypes. Increasingly specific schemas are activated by words and combinations of verbs and agent-, patient-, or instrument-nouns, as well as discourse context (Metusalem et al. 2012). Activated schemas then support a multitude of rapid, parallel stereotypical inferences: At each point, receivers use the most specific inferences to flesh out utterance content, in line with the I-heuristic. The activation processes in semantic memory that support these inferences occur in language-comprehension *and* -production (Pickering

and Garrod 2013, Stephens et al. 2010). Stereotypical enrichment will hence occur not only in interpersonal communication but also in the kind of subvocalized speech involved in philosophical thought (Carruthers 2002).

2. Philosophical Application

We can draw on psycholinguistic research to develop an epistemological profile of this important process, which tells us under what conditions we may (not) trust the stereotypical inferences we automatically make (Weinberg 2015, 2016). This profile can then be used to assess philosophical intuitions and arguments.

2.1. Epistemological profiles

With some *caveats*, stereotypical enrichment is generally reliable. The strength of stereotypical association gradually increases through continued observation of co-occurrence in the physical and discourse environment (seeing more red tomatoes in the supermarket, watching more war movies full of nasty Germans); it gradually decreases, as incompatible observations accumulate (seeing green tomatoes in the fields, meeting friendly Germans) (McRae and Jones 2013, Neely and Kahan 2001). Strength of stereotypical association thus encodes information about co-occurrence frequencies in the subject's physical and discourse environment. To the extent that the physical environment is stable and changes only gradually, and the discourse environment is free from systematic misrepresentation (no war propaganda), probabilistic stereotypical inferences are therefore reasonably accurate. Second, the maxim I-speaker (above) has speakers make deviations from stereotypes explicit, and where contextual cues defeat conclusions of stereotypical inferences, these conclusions typically get swiftly suppressed (Faust and Gernsbacher 1996) or simply decay for wont of reinforcement (Oden and Spira 1983). Such processing largely prevents contextually inappropriate inferences from interfering with utterance comprehension and further reasoning.

We now build up to a set of conditions under which the generally reliable process predictably leads to inappropriate inferences which go through to affect further reasoning, even so (*cf.* Fischer and Engelhardt 2016, under review). This cognitive bias arises from the way in which polysemous words are processed. Most words have more than one meaning or sense (Klein and Murphy 2001). A linguistic stimulus activates all semantic and stereotypical features associated with the expression, in *any* of its meanings or senses, regardless of contextual propriety – e.g. 'mint' activates the probe *candy*, even when used in a different sense (as part of the prime 'All buildings collapsed except the mint') (Simpson and Burgess 1985, Till et al. 1988). The strength of initial activation is ordered by '*salience'*, understood as a function of exposure frequency (how often a language user encounters the word in that sense), modulated by prototypicality (how good examples of the relevant category the word is deemed to stand for, in that sense) (Giora 2003, Giora et al. 2015):[4] The more salient a sense is for a speaker/hearer, the more rapidly and

[4] Exposure frequency cannot be directly measured, but is inferred from occurrence frequencies in corpora, familiarity ratings, or conventionality ratings (Giora 2003). Prototypicality is usually assessed through listing, sentence-completion, or typicality-rating tasks (Battig and Montague 1969, Chang 1986).

strongly the associated situation schema is activated. The more strongly activated a schema is, the longer its activation takes to decay (Farah and McClelland 1991, Loftus 1973) and the more effort is required for its suppression (De Neys et al. 2003, Levy and Anderson 2002). Salience imbalances can therefore lead to an interpretation bias, where utterances employ words in less salient senses or uses – but their interpretation is unduly influenced by the schema associated with the most salient or dominant use.

Such a 'salience bias' (Fischer and Engelhardt, under review) is particularly liable to arise where less salient uses are interpreted with a *Retention/Suppression Strategy* (Giora 2003, 37; henceforward 'Retention Strategy', for short): Where utterances use a polysemous word in a less salient sense, they are often interpreted by retaining the most rapidly activated schema associated with the most salient sense and suppressing the contextually inappropriate component features of the dominant schema (Giora 2003, Giora et al. 2014). This Retention Strategy has been shown to be used in the interpretation of figurative speech (Giora et al. 2007b, 2015), crucially including (non-default) metaphor (Giora et al. 2007a). We would expect it to be involved, e.g. in interpreting metaphorical uses of the verb 'to see': According to a recent corpus study using the *British National Corpus* (Fischer and Engelhardt 2017b), 'see' is used far more frequently in a visual sense (68%) than in an epistemic sense ('I see your point', 12%), a doxastic sense ('as he saw fit', 10%), or a phenomenal sense ('Hallucinating, Macbeth saw a dagger', 1%).[5] The schema associated with the dominant visual sense ('*vision-schema'*) includes agent-features like *S uses her eyes, S looks at X, S knows X is there,* and *S knows what X is* as well as patient-features like *X is in front of S, X is near X, X is there at the same time as S.* To interpret epistemic uses in line with the Retention Strategy, most of these features get suppressed, while retaining the contextually appropriate agent-features *S knows X is there* and *S knows what X is* (yielding the interpretation, 'I know you've got a point and know what it is').[6]

Frequently co-occurring components of a situation schema activate others (Hare et al. 2009, McRae et al. 2005). Where a frequently used word has a dominant sense that is far more salient than all others (like 'see'), the components of the associated schema will frequently co-occur. Initial activation of contextually inappropriate schema components will then not only be strong (due to salience) but also complemented by lateral cross-activation from other schema components. It will then be difficult to suppress only some of the frequently co-occurring components, but not others. Where some of them are retained for utterance interpretation, the others will remain at least partially activated and support contextually inappropriate inferences that go through unsuppressed.

We have thus arrived at a first set of jointly vitiating conditions under which the generally reliable process of stereotypical enrichment is liable to lead to inappropriate conclusions that affect further judgment and reasoning:

---

[5] A production experiment using a sentence-completion task determined the same rank order for prototypicality, with even higher preponderance of the visual sense (Fischer and Engelhardt 2017b). We infer that the visual sense is more salient than the epistemic and doxastic senses and these, in turn, are more salient than the phenomenal sense.

[6] The eye-tracking study reported in Section 5 provides evidence that the Retention Strategy is applied here.

[Hypothesis H] At least <u>when</u>

(i) one sense of a polysemous high-frequency word is much more salient than all others,

(ii) the dominant stereotype (situation schema) is deployed in interpreting utterances involving a less salient use, and

(iii) some, but not all, of the stereotype's frequently co-occurring core components are contextually relevant, <u>then</u>

(1) inappropriate stereotypical inferences licensed only by the dominant sense will be triggered by the less salient use and

(2) these automatic inferences will influence further judgment and reasoning, even when thinkers explicitly know they are inappropriate.

We have thus obtained first components of an epistemological profile of the process of stereotypical enrichment. We now explore how they can be deployed to assess philosophically relevant intuitions and reconstructions of philosophical arguments.

2.2. Philosophical assessments

Arguably, most of the intuitive judgments about, and inferences from, case-descriptions that philosophers make in thought experiments and argument rely on everyday conceptual and linguistic competencies which are also deployed in ordinary discourse (Williamson 2007: 188) – and on the routine cognitive processes underlying these competencies. Insofar as these routine processes are reliable, the philosophically relevant intuitions and inferences should be reliable as well, and the intuitions they generate should have *evidentiary value* (i.e. the fact that thinkers have them, as and when they do, should speak for the intuitions' truth). So just how far are those routine processes reliable?

To address this question, some experimental philosophers have started to develop *'GRECI explanations'* (as we have called them elsewhere, see Fischer and Engelhardt 2016). These explanations trace philosophically relevant intuitions back to cognitive processes which are **g**enerally **re**liable but subject to cognitive biases which generate **c**ognitive **i**llusions under specific conditions. At any rate when produced under these vitiating conditions, intuitions lack evidentiary value. For example, intuitive knowledge attributions elicited through the method of cases have been traced to a mindreading competency (Nagel 2012, Gerken 2017) that has been argued to be generally reliable (Boyd and Nagel 2014), but subject to cognitive biases including a focal bias which may assert itself, for instance, in thought experiments allegedly revealing contrast effects and supporting contrastivism about knowledge (Gerken and Beebe 2016).

Above, we built up towards a GRECI explanation that traces philosophically relevant intuitions not to a domain-specific process like mind-reading, but to a potentially complementary domain-general language comprehension/production process, stereotypical enrichment, which we argued to be generally reliable but subject to cognitive biases including a salience bias. An empirically supported account of this bias will call into question the evidentiary value of intuitions about cases whose descriptions use familiar words in special senses for whose interpretation the dominant sense may be functional. Many philosophical thought experiments involve 'esoteric' cases involving

well-behaved zombies, envatted brains, twin planets, etc. These cases are described with familiar words, given rare new uses, for whose interpretation the dominant sense will typically be functional (in the absence of explicit explanations). To the extent to which the cases deviate from dominant stereotypes (e.g. the 'zombies' behave like us), the dominant stereotypes are then liable to support contextually inappropriate inferences. Our account of salience bias thus supports a variant of the 'esotericity thesis' that intuitions about esoteric cases are less reliable than about 'normal' cases (Weinberg 2007, Williamson 2007). Prior to further psycholinguistic investigation (including examination of the precise salience structure of the relevant words), it provides an undermining defeater of the relevant intuitions (Pollock 1984). Already nascent epistemological profiles which do not (yet) go beyond arguments for general reliability of the target process and the identification of a first set of vitiating conditions thus allow us to assess the evidentiary value of at least some philosophical case-intuitions.

In cognitive psychology, *intuitions* are typically conceptualised as judgments generated by automatic inferences, that is, by autonomous cognitive processes that place low demands on working memory (Evans and Stanovich 2013) and duplicate inferences with heuristic rules (Kahneman and Frederick 2005). Such processes may issue either in explicit judgments ('intuitions') or conclusions that are tacitly presupposed in further cognition (judgment and reasoning). 'Experimental philosophy 2.0' seeks to extend epistemological investigation from intuitions to arguments (Nado 2016) and we now focus on automatic inferences driving philosophical argument.

Our example is taken from the philosophy of perception, where philosophers wishing to merely describe perceivers' subjective experience systematically use familiar appearance- and perception-verbs in a rarefied 'phenomenal' sense, which lacks existential, factive and spatial implications (e.g. Ayer 1956: 90, Fish 2010: 6, Jackson 1977: 33-49, Macpherson 2013: 5; *cf.* Chisholm 1957: 44-48). We submit that it satisfies condition (i)-(iii) from Hypothesis H: At any rate for 'see', we have shown (i) that the verb has a clearly dominant (visual) sense and that the phenomenal sense is the least salient sense (above, Fn.5). We suggest that (ii) the dominant word schema is retained and deployed to interpret the latter (*cf.* Giora 2003, Giora et al. 2014): A situation-model that instantiates the dominant schema with specific patient-role fillers is constructed. This model contains a set of phenomenal features as a component, and these features are attributed to the target experience, in a variant of the common 'feature transfer' approach of metaphor interpretation (Bortfeld and McGlone 2001, Ortony 1993). However, (iii) what it is like to see something is strongly associated with the other features of the schema associated with the dominant use of 'see', as evidenced by embodied cognition effects associated with visual metaphors (Lakoff 2012, Landau et al. 2010). Accordingly, it is hard to retain only the phenomenal component and suppress the schema's other core components. Hypothesis H therefore predicts that uses of the rarefied phenomenal sense will prompt contextually inappropriate (e.g. existential and spatial) inferences supported – only – by the dominant visual sense of the verb.

The special phenomenal sense is then used to talk about unusual cases, like hallucinations. Where thinkers know little about the relevant cases (e.g. hallucinations),

conclusions are yet less likely to be suppressed through integration with background knowledge (*cf.* Metusalem et al. 2012, Fischer and Engelhardt 2017a), and yet more likely to affect further cognition (judgment and reasoning). We therefore regard it as particularly likely that 'arguments from hallucination' will involve such contextually inappropriate stereotypical inferences.

In their traditional version, these arguments argue for the existence of mind-dependent objects of sense-perception ('sense-data'), which separate subjects from any physical objects in the environment (Ayer 1956: 90, Fish 2010: 12-15, Macpherson 2013: 12-13, Smith 2002: 194-197, *cf.* Crane and French 2015: sec.3.1). Here is a classic statement that explicitly marks the special phenomenal sense used:

> 'Let us take as an example Macbeth's visionary dagger [...] There is an obvious [perceptual] sense in which Macbeth did not see the dagger; he did not see the dagger for the sufficient reason that there was no dagger there for him to see. There is another [viz., phenomenal] sense, however, in which it may quite properly be said that he did see a dagger; to say that he saw a dagger is quite a natural way of describing his experience. But still not a real dagger; not a physical object... If we are to say that he saw anything, it must have been something that was accessible to him alone… a sense-datum.' (Ayer 1956: 90).

The second half of the argument then generalises from this special case (hallucination) to all cases of visual perception: Since subjectively indistinguishable experiences (supposedly) must be mental states of the same type, and mental states of the same type must have objects of the same kind (mind-dependent or –independent), actual perceptual experiences must have the same objects of awareness as possible hallucinatory experiences that are subjectively indistinguishable from them.

But note the persuasive fallacy in the first half: Macbeth is meant to have an experience just like that of seeing a physical dagger. In the phenomenal sense (where 'S sees an F' means 'S has an experience like that /as of seeing an F'), he therefore *can* be said to 'see a physical dagger' (his visual experience is, by assumption, just like that of seeing a physical dagger) – while he cannot be said, e.g. to see a translucent non-physical dagger (his experience is not like that). In the quoted passage, the move from

(1) 'Macbeth saw a dagger' (in the phenomenal sense)

to 'but still not a real dagger' is hence fallacious. We suggest the argument is driven by inappropriate inferences from the phenomenal use of 'see' to conclusions that typically remain implicit, but are presupposed in further reasoning, e.g.

(2) There then was something that Macbeth saw. – But, by assumption:
(3) 'There then was no physical object for Macbeth to see.' By (2) and (3):
(4) 'There then was a non-physical object that Macbeth saw'.

The inference from (1) to (2) would be licensed by the dominant visual sense of 'see', but not by the contextually relevant phenomenal sense which lacks factive implications and creates an intensional context not admitting of quantification (Forbes 2013). The same conclusion can be reached also by spatial inferences, also supported only by the visual sense (Fischer and Engelhardt 2017b). Given explicit marking of the different uses in the quoted passage, more must be involved than a simple error of using the wrong sense of 'see'.

Plausible principles of charity limit the extent of irrationality and conceptual or linguistic incompetence we may attribute to competent thinkers (Adler 1994, Lewinski 2012). Empirical explanations of why competent thinkers commit fallacies are then required to validate reconstructions (Thagard and Nisbett 1983). Our proposed account of salience bias can explain why competent speakers should make contextually inappropriate stereotypical inferences from 'see' – and other verbs, used in other statements of the argument (Fischer and Engelhardt, under review). But is this account correct? Do competent language users make inferences licensed only by the dominant use of a polysemous verb also from rarefied special uses, as posited by Hypothesis H?

2.3. Pre-study

We used a battery of complementary psycholinguistic methods to examine this hypothesis about automatic inferences which speakers/hearers are not aware of making. The hypothesis boldly claims that under certain condition (i-iii above), competent language users will go along with contextually inappropriate stereotypical inferences *even when they know the inferences at issue to be inappropriate*. In a pre-study (see Fischer and Engelhardt, under review), we identified potentially relevant inferences: Undergraduate participants rated spatial inferences from visual, epistemic, doxastic, and phenomenal uses of 'see', as well as from visual and epistemic uses of 'aware'. For example:

	[*visual 'see'*]	[*epistemic 'see'*]
[Premise]	Mona sees the car on the road.	Josh sees the issues in play.
[Conclusion]	The car on the road is in front of Mona.	The issues in play are in front of Josh.

Table 1. Spatial inferences from different uses of 'see'.

On a 5-point Likert scale, participants indicated their confidence that 'in situations where the first sentence [premise] is true also the second sentence [conclusion] will typically be true'. They were very confident that spatial inferences from visual uses typically lead to true conclusions (mean rating 4.7 for 'see', and 3.7 for 'aware', both significantly above neutral mid-point '3'). They were also very confident that spatial inferences from the other uses examined typically lead to conclusions that fail to be true (mean ratings significantly below 3). They were most confident that spatial inferences from epistemic uses are inappropriate (lead to conclusions that fail to be true) (mean rating 1.58 for 'see' and 1.53 for 'aware'). To get clear on whether competent language users make contextually inappropriate stereotypical inferences under the conditions (i)-

(iii) identified by our Hypothesis H, we therefore examined whether speakers/hearers make and deploy spatial inferences from epistemic uses of 'see', which they demonstrably know to be inappropriate.[7] We accordingly examined the verb-specific hypotheses:

H_1 Competent speakers infer spatial patient-properties (*X is in front of S*) from visual *and epistemic* uses of 'S sees X'.

H_2 Conclusions from *all* these inferences will be deployed in subsequent cognitive processing, regardless of contextual (im)propriety.

3. Plausibility ratings

3.1 Approach and predictions

Plausibility ratings offer a convenient first means for following up hypotheses about comprehension inferences from specific words that affect further cognitive processing. They thus allow us to examine at any rate the conjunction of H_1 and H_2. Participants hear or read sentences like the following, and rate their plausibility on a Likert scale:

1a. Matt sees the spot on the wall facing him. (*'s-consistent'*)

2a. Chuck sees the spot on the wall behind him. ('*s-inconsistent'*)

In these sentences, the expression of interest is followed by a sequel that is either consistent with the hypothesised stereotypical inference ('s-consistent') or inconsistent with it ('s-inconsistent'). Our items have post-verbal contexts that are either consistent or inconsistent with the hypothesised inference from 'S sees X' to 'X is in front of S'. If this inference is made, and not swiftly suppressed, then the clash with the sequel will render s-inconsistent items (like 2a) less plausible than s-consistent items (like 1a).

This prediction holds both on a content- and an experience-based approach to metacognitive judgments. If the plausibility judgment is based on cognitive engagement with the content and an assessment of its probability, the clash of the s-inconsistent sequel with the conclusion of the probabilistic inference (e.g. in 2a) will make its truth less probable than that of its s-consistent counterpart (1a). According to the experience-based approach to metacognitive judgments (Koriat 2007), the subjective plausibility of a judgment results not from cognitive engagement with its content but from features of the underlying cognitive processes. Fluency or level of effort serves as a cue for a wide range of metacognitive judgments, including plausibility assessments (for a review, see Alter and Oppenheimer 2009). Perceived inconsistencies (as in 2a) reduce the degree of 'fluency' or effortlessness of the comprehension process (Carpenter and Just 1977), which in turn reduces subjective plausibility (Thompson et al. 2011). Either way, lower plausibility ratings for s-inconsistent sentences than for their s-consistent counterparts would provide evidence for automatic spatial inferences.

---

[7] If correct assessment does not prevent inappropriate automatic inference to conclusions presupposed in further reasoning, this will provide some support for inferences from findings about undergraduates to conclusions about expert philosophers, in the light of the 'expertise objection' (review: Machery 2015).

To show that the inferences of interest are supported by features of the verb (e.g. stereotypical features), we manipulate not only the post-verbal context but also the verb and replace 'see' in half the items by a contrast verb less strongly associated with spatial patient features. We employ 'is aware of', which is ordinarily used in an epistemic sense, to attribute knowledge that may, but need not, be acquired through the five senses (*MEDAL*, *WordNet*).[8] A prior production experiment with a sentence-completion task showed that this verb is paired about half the time with visual objects as patients, which agents would be aware of in virtue of looking at them – whereas 'see'-stems are provided with completions that give the verb a visual sense, over 93% of the time (Fischer and Engelhardt 2017b). We infer that 'aware' is less strongly associated with the vision-schema, and spatial patient-properties, than 'see', and include items like

1b. Matt is aware of the spot on the wall facing him.

2b. Chuck is aware of the spot on the wall behind him.

Again, content- and experience-based accounts support the same prediction from our hypotheses: The weaker association of 'aware' (than 'see') with spatial features translates into a weaker probabilistic inference (It's less probable that the patient is in front of the agent), making it more probable that the sentence is true (the agent knew all along, from previous observation or testimony, that there's a spot on the wall). Less strongly supported inferences are also easier to suppress, leading to less disfluency. Either way, s-inconsistent 'aware'-sentences (like 2b) will be rated more plausible than their 'see'-counterparts (like 2a). If so, this will provide evidence that the inferences of interest are supported by features of the verb.

To examine whether spatial inferences are also, inappropriately, made from epistemic uses of the verb, we finally manipulate the object. In the absence of contextual cues, concrete patient-nouns ('picture', 'car') invite visual interpretations of 'see'. By contrast, epistemic readings are invited by abstract patient-nouns ('challenges', 'opportunities', henceforth 'epistemic objects', for convenience), whose referents typically cannot be literally 'seen', but known. We use items like:

3a. Joe sees the problems that lie ahead.

3b. Joe is aware of the problems that lie ahead.

4a. Jack sees the problems he left behind.

4b. Jack is aware of the problems he left behind.

In principle, perfectly plausible interpretations are readily available for s-inconsistent sentences with epistemic objects (like 4): We can complement a purely epistemic interpretation of 'see' or 'aware' with a metaphorical interpretation of the sequel (before subject=future, behind subject=past; hence 'Jack knows what problems he had in the past'). But incompletely suppressed spatial inferences from 'see' will prevent such purely metaphorical interpretation: Though conventional, the present space-time

[8] https://www.macmillandictionary.com/dictionary/british/aware, http://wordnetweb.princeton.edu/perl/webwn

metaphors give rise to embodied cognition effects (Boroditsky and Ramscar 2002, Bottini et al. 2015) and support spatial reasoning about temporal relations (Casasanto and Boroditsky 2008, Gentner et al. 2002). We infer that these metaphorical sequels will activate spatial schemas that place objects in front of, or behind a forward/future-facing subject (Gentner et al. 2002). These schemas will be retained during comprehension of these sequels (Giora 2003, Giora and Fein 1999), and facilitate spatial reasoning from them (Casasanto and Boroditsky 2008). If spatial inferences from the prior verb are made, their conclusions will therefore engage such reasoning, and be felt to clash with s-inconsistent sequels ('...left behind'). These perceived clashes will engender comprehension difficulties and render s-inconsistent 'see'-items (like 4a) less plausible than s-consistent counterparts (like 3a).

'Aware'-items will elicit different responses: Even if the association of 'is aware of' with the vision-schema is weaker, the combination of the verb with a *visual* object-noun will activate the schema (*cf.* Bicknell et al. 2010) and prompt spatial inferences. However, their conclusions will, where contextually necessary, be easier to suppress than those from 'see' (above). Therefore s-inconsistent 'aware'-sentences with visual objects (like 2b) will be deemed less plausible than their s-consistent counterparts (like 1b), but still more plausible than otherwise identical 'see'-sentences (like 2a). By contrast, if the dominant use of 'is aware of' is epistemic, and any spatial conclusion is inappropriate where it goes with abstract epistemic objects, such conclusions should be swiftly and completely suppressed in interpreting 'aware'-sentences with *epistemic* objects. For such sentences, we therefore expect spatial inferences from 'aware' will not interfere with subsequent plausibility judgments, so that the context-manipulation will not affect plausibility ratings – which should hence again be higher than for 'see'-counterparts.

We thus derive two key predictions from our hypotheses:

[Plausibility-1] s-inconsistent 'see'-sentences, both with visual *and* with epistemic objects, will be deemed less plausible than their s-consistent counterparts.

[Plausibility-2] s-inconsistent 'see'-sentences, both with visual *and* with epistemic objects, will be rated less plausible than their 'aware'-counterparts.

To sum up, we can test the hypotheses that competent language users make inappropriate stereotypical inferences from a specific verb (here: 'see'), which influence further cognition, by using a plausibility-rating task and a 2 × 2 × 2 (context [s-consistent/s-inconsistent] × verb [see/aware] × object [visual/epistemic]) design, where all variables are manipulated within-subject.

3.2 Excluding confounds

This design helps to exclude most potential confounds. The verb-manipulation helps us exclude spatial inferences from other parts of the sentence as drivers of plausibility judgments. First, patient nouns might be associated with a specific spatial orientation towards agents (e.g. 'challenges' are typically said to be 'ahead'). However, if patient-nouns were the prime source of spatial inferences, they should reduce the plausibility of

s-inconsistent ‘see’- and ‘aware’-sentences in the same way and falsify prediction Plausibility-2.[9]

Second, in items with epistemic objects, the spatial inference might be triggered by the spatial time metaphor used by the post-verbal sequel. This metaphor activates a schema centred on a subject spatially oriented to look at things in front of her (Gentner et al. 2002). One might therefore argue that *S looks at things in front of S* is a component of this schema, and that s-inconsistent sentences like ‘Jack sees the problems he left behind’ will seem implausible because they clash with this component. However, in this case, both ‘see’- and ‘aware’-items with epistemic objects should be sensitive to the context-manipulation. If this manipulation affects the plausibility only of ‘see’, but not ‘aware’-items, we can exclude this potential confound.

Third, lower mean plausibility ratings for s-inconsistent ‘see’-sentences might be driven by existential or factive inferences from the verb, which are appropriate both with visual and epistemic objects (You can only see my point if I have one), as these too are cancelled by half our s-inconsistent sequels with epistemic objects, like ‘the problems he left behind’ (which implies the problems no longer exist for the agent), though not the others (like ‘the possibilities from which she has turned away’). However, existential and factive implications are shared by both ‘see’ and ‘aware’ (You can only be aware of a problem that actually exists), so that the plausibility of epistemic ‘aware’-items should again be affected in the same way by the context manipulation.

Finally, plausibility judgments can be affected by word frequency, as more frequently encountered words are easier to process (Oppenheimer 2006) and higher fluency may serve as metacognitive cue anchoring plausibility judgments (Alter and Oppenheimer 2009). While this is no major issue in the present studies (since we predict lower plausibility rankings for items using the more frequent ‘see’, and mean frequencies for our forward and backward terms are very similar) we can assess the extent to which frequency influences participants’ ratings by constructing ‘frequency-congruent’ filler items where the sentence with the more frequent verb is also more consistent with its associated stereotype, and ‘frequency-reversed’ filler items, where word-frequency and stereotype-consistency work in opposite directions. If participants make judgments predominantly in line with stereotype-consistency, and no fewer such judgments about frequency-reversed than frequency-congruent items, their plausibility judgments are unlikely to be influenced by frequency, and we can assess whether their ratings (still) bear out our predictions (Fischer and Engelhardt 2016).

Since philosophical thought takes place in both speech (oral debate) and text (reading and writing), we investigated inferences in both modalities. We employed the plausibility-rating paradigm described in two studies, in which participants heard and read

---

[9] We investigated this possibility, even so: In a norming study, participants considered the patient nouns used in Study 1 (Section 4), in ‘aware’-contexts (e.g. ‘Joe is aware of the problems’, etc.), and rated them on whether what they stand for is typically ‘ahead’ (=1) or ‘behind’ (=-1) the agent, or one ‘cannot tell’ (=0). Mean ratings >0 indicate forward bias. While the overall mean for our visual nouns did not differ significantly from 0, that for our epistemic objects did. Precisely half had mean ratings significantly >0, the other half had means not significantly different from 0 (neutral). In the main study, our two key predictions were borne out by ratings for items with forward-oriented *and* ‘neutral’ patient nouns.

the items, respectively, and the task was combined with appropriate online measures. We provide results for each after outlining the relevant further methods.

4. Pupillometry

For our purposes, the major shortcoming of offline (outcome) measures like plausibility ratings is that they allow us to examine only hypotheses about conclusions which are automatically inferred *and* maintained. To examine separately which inferences are initially made automatically at the verb of interest (as per H_1) and to what extent their conclusions are subsequently suppressed or maintained (as per H_2), we can combine plausibility ratings with pupillometry or other eye tracking measures. Since we report the pupillometry study elsewhere (Fischer and Engelhardt, under review; see Fischer and Engelhardt 2017a for a pilot), we here focus on explaining the approach and give only an executive summary of methods and results, before reporting a fresh reading time study (Section 5).

4.1 Approach and predictions

Pupil dilations offer a window into preconscious automatic processing (reviews: Laeng et al. 2012, Sirois and Brisson 2014). Our pupils dilate when we are emotionally or cognitively aroused or expend cognitive effort. Pupil responses to task demands (as opposed to, say, changes in lighting) are highly correlated with neural activity in the *locus coeruleus*, a key node within neural circuitry that controls the muscles of the iris (Samuels and Szabadi 2008) and mediates the functional integration of the whole attentional brain system (Corbetta et al. 2008). Pupil diameter reliably increases with the 'intensity' of attention (Kahneman 1973) or cognitive load (the extent to which cognitive resources are mobilised to address a task). These pupil responses are spontaneous and impossible to suppress at will (Loewenfeld 1993); they are triggered also by subliminally presented stimuli the subject is not aware of (Bijleveld et al. 2009), and regularly commence well before any conscious task response. They thus allow us to gauge allocation of cognitive resources at pre-conscious stages of processing (Laeng et al. 2012).

In language processing, difficulties which require cognitive resources to overcome arise from several sources. While psycholinguists have only recently begun to take up pupillometry on a larger scale, pupil responses have been found sensitive to syntactic complexity and sentence length (Piquado et al. 2010) and differences in the intelligibility of speech due to interfering noise (Zekveld and Kramer 2014), where dilations peak at medium levels of interference, suggesting less resources are allocated when the task becomes too difficult. The level of difficulty is generally also dependent upon the predictability of new text in the light of old: Processing is facilitated by activation of subsequent concepts by previous words, through associative priming (based on co-occurrence of words) or semantic priming (based on activation of schemas and semantic knowledge, more generally). Accordingly, pupil responses have been found responsive to '*surprisal'*, that is, the predictability of the next word in a sentence, given its previous words, as estimated, e.g. by recurrent neural networks, on the basis of co-occurrence frequencies (Frank and Thompson 2012). By contrast, where new text violates

expectations and clashes with the conclusions of schema-based inferences, suppression is required and costs effort (Faust and Gernsbacher 1996). Accordingly, pupillometry has documented dilations in response to violations of scripts (social event/action schemas) (Raisig et al. 2012).[10]

Our study used pupillometry to garner evidence of – inappropriate – stereotypical inferences supported by event schemas. While pupil dilations are initiated rapidly, the pupil takes over one second to dilate to its maximum size, after the point of difficulty (Engelhardt et al. 2010). Since it does not respond at uniform speed to all kinds of difficulties, it may, in this period, be influenced also by difficulties preceding or following the difficulty of interest. To minimise such influence, we created the difficulty at the end of our sentence items and compare mean pupil sizes after sentence offset with mean sizes in the previous time window (rather than considering time course). In our items, the difficulty of interest arises from a clash between the last part of the sentence and inferences from the prior verb (or verb and object).

Our hypothesis H_1 thus predicts that

[Prediction Pupil] s-inconsistent, but not s-consistent, 'see'-sentences with visual *and* with epistemic objects will prompt pupil dilations in the second after sentence offset.

Since the combination of 'is aware of' with a visual object will activate the vision-schema at the earliest possible moment (*cf.* Bicknell et al 2010), i.e. at the object-noun, we also expect significant dilations in the sentence offset window for s-inconsistent 'aware' sentences with visual objects, but not with epistemic objects.

Since pupil dilations are sensitive to effort expended at pre-conscious stages of processing, they can provide evidence of inferences, even where conclusions get swiftly suppressed and fail to influence subsequent judgments. This can happen where conclusions conflict with background knowledge or beliefs that get swiftly activated in language comprehension (Metusalem et al. 2012). In a study we ran jointly with the present experiment (Fischer and Engelhardt 2017a, Exp.2), we found participants' plausibility judgments were highly sensitive to their content-related background beliefs: Two groups held opposing views on the issue at hand (whether homophobic attitudes are pathological). Pupillometry results suggested they made the same stereotypical inferences from the expression of interest ('S is homophobic'), to conclusions (*S is mentally ill*) consistent with the background beliefs of one group of participants, but not another. Even so, the groups proceeded to give opposite plausibility ratings for sentences with sequels consistent and inconsistent, respectively, with the conclusions (e.g. 'Tim is homophobic. He is mentally ill' vs. 'Joe is homophobic. He is mentally healthy'). Where initial conclusions were inconsistent with their background beliefs, participants suppressed them sufficiently swiftly and comprehensively to prevent them from influencing plausibility judgments.

[10] Recent evidence suggests pupil responses may index conflict monitoring yet more reliably than cognitive effort (van Steenbergen and Band 2013). This would strengthen the case for using pupillometry to investigate automatic inferences through pupil responses to subsequent cancellation phrases.

In our paradigm, we therefore take pupillometry and subsequent plausibility ratings to measure different things: Pupillometry picks up inferences automatically made in incremental language comprehension and allows us to examine hypothesis H_1. Plausibility ratings measure the extent to which inferences are successfully suppressed or influence subsequent cognition and let us examine hypothesis H_2.

The pupil is far more responsive to luminance variations than to changes in cognitive load (Beatty et al. 2000). Since the presentation of reading items on ordinary-sized computer screens involves luminance differences as eyes move from the beginning of the sentence (when the visual field extends beyond the screen) to the centre of the screen, only few pupillometric investigations into language processing employ reading tasks (e.g. Frank and Thompson 2012, Raisig et al. 2012). Our study employs a listening task, which allows participants to fixate a fixation cross in the middle of a computer screen, throughout the trial.

4.2 Methods

Our thirty-eight participants were undergraduate students. All were native speakers of English. They heard sentences including 48 critical items, viz., 6 for each of the eight conditions (examples 1a – 4b, Section 3.1). Items in each category alternated post-verbal contexts (e.g. 's/he left behind' and 's/he had turned away from', for s-inconsistent items with epistemic objects) and employed the same epistemic patient nouns as the pre-study (Section 2.3).

We used a 2 × 2 × 2 (context [s-consistent/s-inconsistent] × verb [see/aware] × object [visual/epistemic]) design and manipulated all variables within subject. Participants were seated at the eye tracker, given a set of verbal task instructions, and placed their chins on a chinrest. After a calibration procedure, they completed practice and experimental trials. On each trial, a fixation cross appeared for 1500 ms prior to sentence onset. The pre-recorded sentence was played out on the computer speakers, and after sentence offset the fixation cross remained on the screen for 1000 ms. After the cross disappeared, a plausibility rating prompt appeared, and participants rated sentences' plausibility from 1 to 5, using the corresponding key on the keyboard. Mean pupil diameter was measured with an SR Research Eyelink 1000 in time windows including the second half of the sentence and the offset period.[11] We baseline corrected the pupil diameter based on the preceding time window: We divided the mean size of the pupil during offset by the mean size during the second half of the sentence, for each condition. This allowed us to assess whether the pupil size was changing between time windows. To do so, we conducted one-sample t-tests with a test value of 1. A value of 1 would indicate that mean pupil diameter remained the same.

4.3 Results and discussion

All our predictions were confirmed. Pupil results are shown in Figure 1.

[11] For technical specifications of this frequently used device, see https://www.sr-research.com/wp-content/uploads/2017/11/eyelink-1000-plus-specifications.pdf

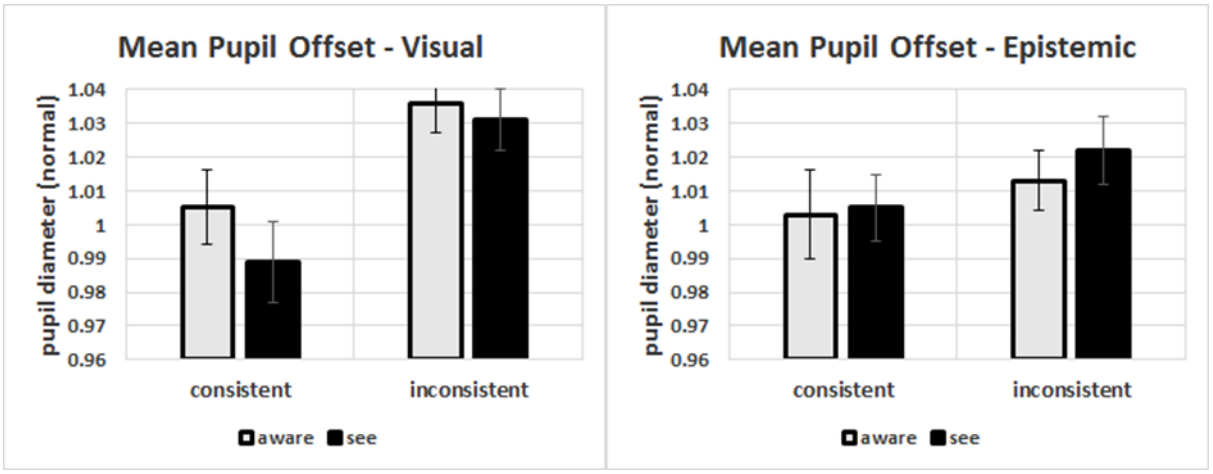

Figure 1. Baseline adjusted pupil diameter in 1000ms time window following sentence offset. Error bars show the standard error of the mean.

The s-inconsistent items resulted in larger pupil diameters. Crucially, participants' pupil size significantly increased after hearing s-inconsistent 'see'-sentences with visual *and* with epistemic objects. As further expected, there was also an increased pupil size after hearing s-inconsistent 'aware' sentences with visual objects, but no significant increases in the other conditions. However, while remaining shy of significance, pupil dilations after s-inconsistent 'aware' sentences with epistemic objects, clearly fell between the dilations observed in the other conditions with epistemic objects. We interpret this unexpected finding as evidence that the weak association of 'aware' with the vision-schema still supports initial spatial inferences.

Plausibility results are given in Figure 2.

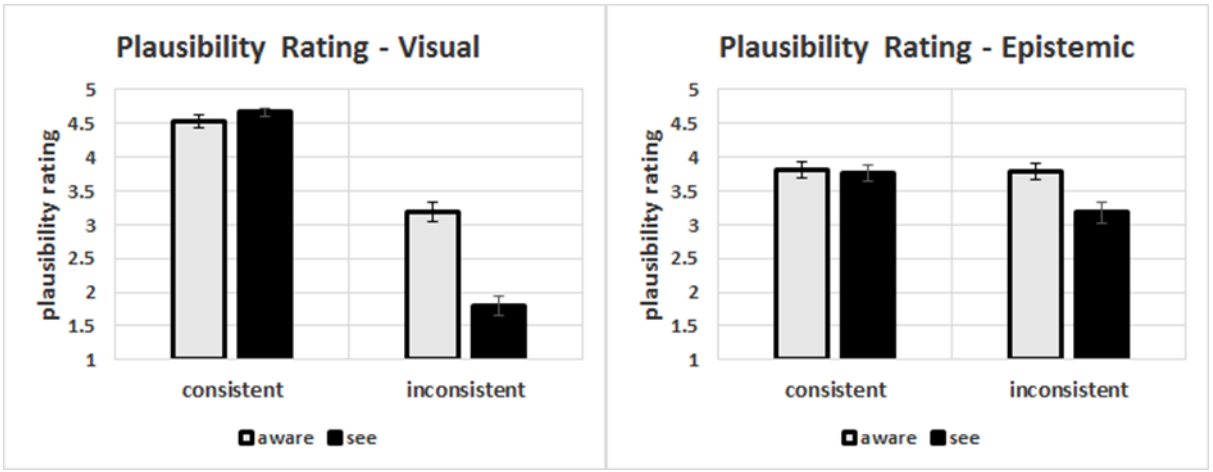

Figure 2. Mean plausibility ratings for each of the eight conditions in the pupillometry study. Error bars show the standard error of the mean.

S-inconsistent 'see' sentences with visual *and* with epistemic objects were deemed less plausible than s-consistent counterparts (as per prediction Plausibility 1) and s-inconsistent 'aware'-counterparts (as per Plausibility 2). Items in all s-consistent conditions were judged distinctly plausible (mean ratings significantly above mid-point 3), as were s-inconsistent 'aware'-items with epistemic objects, while s-inconsistent 'see'-items with visual objects were judged distinctly implausible (mean ratings below

3), and s-inconsistent ‘see’-items with epistemic objects were deemed neutral (not significantly different from 3) (as were s-inconsistent ‘aware’-items with visual objects). Predicted plausibility differences thus materialised as categorial differences.

Plausibility results cohere with pupillometry findings: Precisely in the three conditions with significant pupil dilations participants refrained from rating sentences as ‘plausible’. The absence of more precise mirroring is consistent with ur view that pupil dilations and plausibility ratings measure different things (Section 4.1): We interpret the observed pupil dilations as evidence of automatic inferences – including contextually inappropriate spatial inferences from epistemic uses of ‘see’. We regard subsequent plausibility ratings as evidence of the extent to which suppression is successful. A purely epistemic interpretation of s-inconsistent items with epistemic objects (where ‘Jack sees / is aware of / the problems he left behind’ comes down to *Jack knows what problems he had in the past*, see Section 3.1), renders these items at least as plausible as their s-consistent counterparts (which then come down to, e.g. *Joe knows what problems he will have in the future* – that is notoriously hard to predict). We observe almost identical plausibility ratings for s-inconsistent and s-consistent ‘aware’-sentences with epistemic objects. If the vision schema was deployed in interpreting epistemic uses of ‘aware’, the initial spatial inferences unexpectedly suggested by our pupillometry findings were suppressed with complete success, and a purely epistemic interpretation attained, before plausibility judgments were made.

This is different for items with ‘see’, which is more strongly associated with the vision-schema: Lower plausibility ratings for s-inconsistent ‘see’-sentences with epistemic objects than for s-consistent counterparts and for analogous ‘aware’-sentences suggest that inappropriate spatial conclusions inferred from epistemic uses of ‘see’ were not completely suppressed and prevented purely epistemic interpretation. Together with our pre-study, present findings provide evidence of contextually inappropriate inferences (from ‘see’) competent hearers make and presuppose in further cognition, despite knowing they are inappropriate.

5. Eye tracking: Fixation times

Philosophical thought takes place in reading and writing as well as in oral debate. We therefore followed up pupillometric investigation of automatic inferences in speech comprehension with a study that combined plausibility ratings with tracking of eye movements to examine H_1 and H_2 as hypotheses about automatic inferences in reading. This paradigm has the advantage of allowing us to localise the source of comprehension difficulties more precisely, as eye movements respond to difficulties more quickly than pupil size and display more intricate response patterns across different (early and late) eye movement measures. By exploiting these advantages, we can also close the remaining gap in the argument initially motivating our verb-specific hypotheses H_1 and H_2: We can now examine our suggestion that epistemic uses of ‘see’ are interpreted with the Retention Strategy (Section 2.1).

5.1. Approach and predictions

Contrary to a common folk conception, reading is *not* a sequential process where each word in a sentence is read, one after the other, as they appear in the text, at roughly the same pace. Instead, the eye moves in stops (fixations on words) and starts (saccades). Readers tend to fixate most, but not all words, as their eyes move forward (skipping the words easiest to predict from the context) *and* backwards ('regressions' at points of difficulty). According to the 'good enough processing' approach that informs much eye tracking research on reading (Ferreira et al. 2002, Ferreira and Patson 2007, *cf.* Frazier and Fodor 1978) and is consistent with broader trends in cognitive science (Ferreira and Lowder 2016), hearers/readers immediately construct local interpretations over small numbers of adjacent words; if the task at hand demands it, and only then, they subsequently integrate these local interpretations into more comprehensive interpretations of longer sentences and passages, which take long-distance dependencies into account (Swets et al. 2008). In reading, we thus need to recognise words and integrate them into local and more comprehensive interpretations. Difficulties at these different stages manifest themselves in different eye-tracking measures (*cf.* Section 1.2).

The difficulty of word recognition depends mainly on the word's frequency, length, and predictability in the (local) context (Clifton et al. 2016, Rayner 1998). It is reflected in first-pass reading (fixation) times. A backward eye movement (regression) upon first fixation may indicate difficulty in integrating the word into a local interpretation. The regression path duration (sum of [1] all fixations on a word or in a region before moving to the right, plus [2] all fixations made during regressions in this period) reflects the effort required to overcome this difficulty. By contrast, difficulties in integrating the local interpretation of a sentence region into a more comprehensive interpretation will show up only in increased second-pass or total reading times for the region, and a higher number of saccades from it to other text (Clifton et al. 2007, Rayner et al. 2004). Difficulties arising from one sentence region may lead to longer total reading times for the next (n+1) region (Rayner et al. 2004).

We wish to examine, first, our hypothesis that epistemic uses of 'see' are interpreted by retaining the dominant vision-schema and suppressing its contextually irrelevant components. Whenever an ambiguous word with a clearly dominant meaning is used in a less salient sense and disambiguated by immediate post-verbal context (still considered in constructing local interpretations), this increases first-pass reading times and regression-path durations on the disambiguating region, as well as the number of regressions from it (Sereno et al. 2006). Use of the Retention Strategy (Giora et al. 2014), however, implies more sustained suppression effort (Faust and Gernsbacher 1996) and should translate also into longer total reading times for the disambiguating region which are not entirely driven by longer first-pass reading times. In other words, this sustained effort should translate into longer total and second-pass (= total minus first-pass) reading times for the disambiguating region. In our critical items, 'see' is disambiguated by the visual vs epistemic object-noun immediately following it. Higher second-pass and total reading times can also be driven by regressions from the following post-patient context,

prompted by integration difficulties. In our critical items, such difficulties arise from s-inconsistent, but not s-consistent contexts. Our hypothesis thus motivates:

[Prediction EM1] First-pass, second-pass, and total reading times for the object region will be longer for 'see' sentences with epistemic than with visual objects, across all 'see'-items *and* specifically in s-consistent sentences.

The interpretation of epistemic uses of 'aware' will show a partially distinct pattern. Epistemic objects are more abstract than visual objects, and this fact alone increases early processing effort (Binder et al. 2005) and first-pass reading times (Schwanenflugel and Shoben 1983). But differences should show up in later processing stages. According to our initial assumptions, the interpretation of epistemic uses of 'is aware of' involves discarding the weakly associated vision-schema in favour of a different, purely epistemic, situation schema. Our pupillometry findings cast first doubt on this assumption and suggest that the vision-schema may be deployed to interpret epistemic uses also of 'is aware of'. However, while salient, the visual use arguably is not as clearly dominant for 'aware' as the visual sense is for 'see'.[12] The resulting weaker association of the vision schema with 'aware' than 'see' should then translate into less effort being required to suppress contextually inappropriate schema components, and more comprehensive suppression success. The latter success would manifest in plausibility ratings, the former effort in 'late' eye movement measures. Our initial *and* our modified assumptions about 'aware' thus both motivate

[Prediction EM2] Second-pass and total reading times on epistemic objects will be longer for 'see' sentences than for their 'aware'-counterparts.

These measures alone will not allow us to adjudicate between assumptions concerning 'aware', but can provide evidence of the vision-schema's retention in interpreting epistemic uses of 'see'.

Second, we wish to examine our key hypothesis H_1 that competent speakers infer spatial patient-properties (*X is in front of S*) from visual *and epistemic* uses of 'S sees X'. To do so, we construct sentences where visual and epistemic objects, respectively, are followed by sequels that are s-consistent ('that lie ahead of him') or s-inconsistent ('that lie behind him'). We assume the previous text (e.g. 'Joe sees the problems') constitutes a local context, so the clash between the spatial inference and the s-inconsistent sequel will arise at the stage of integrating the local into larger interpretations (Ferreira et al. 2002). Difficulties at this stage show up in 'late' measures. From H_1 we therefore infer

[Prediction EM3] In 'see'-sentences with visual *and* epistemic objects, total reading times will be longer for s-inconsistent post-object contexts than for s-consistent counterparts.

[12] While data on occurrence frequencies for different uses of 'aware' remains to be collected, participants in a production study used visual patient nouns about half the time to complete sentence-stems with 'aware', while providing completions resulting in a visual use of 'see' 94% of the time (Fischer and Engelhardt 2017b).

The relevant clashes can also show up through increased regressions. However, to keep our items as similar as possible to the pupillometry study, we placed these post-object contexts at the end of the sentence, where regressions routinely occur as part of a 'wrap-up' process, anyway (Rayner et al. 2000). We therefore make no predictions about regressions.

Comparing total reading times for post-object contexts (see Table 2) in 'see'- and 'aware'-sentences with epistemic objects can help us adjudicate between assumptions about the processing of epistemic uses of 'aware': If the vision-schema is discarded for a dominant epistemic schema, processing effort should focus on the epistemic object, rather than the post-object context, and the consistency-manipulation should affect total reading times for the context region less than when the vision-schema is retained for interpreting the utterance. We assume the latter holds for 'see'-sentences. Our initial assumption that the vision-schema is not retained to interpret epistemic uses of 'is aware of' then implies

[Prediction EM4] Total reading times of s-inconsistent post-object contexts will be longer for 'see'-sentences with epistemic objects than for their 'aware'-counterparts.

Disconfirmation of this prediction would favour our modified assumption that the vision-schema is retained in interpreting epistemic uses also of 'aware'.

Our final hypothesis H_2 is, to repeat, that, regardless of contextual (im)propriety, spatial conclusions from both visual *and* epistemic uses of 'see' will be deployed in subsequent cognitive processing beyond utterance comprehension. This hypothesis is again assessed through subsequent plausibility-ratings. *Mutatis mutandis*, the above reasoning (Section 3.1) continues to apply and motivate two predictions (to repeat from above):

[Plausibility-1] S-inconsistent 'see'-sentences, both with visual *and* with epistemic objects, will be deemed less plausible than their s-consistent counterparts.

[Plausibility-2] S-inconsistent 'see'-sentences, both with visual *and* with epistemic objects, will be rated less plausible than their 'aware'-counterparts.

5.2. Methods

Participants. Thirty-six undergraduate psychology students from the University of East Anglia, thirty women, six men, ranging in age from 18 to 26 years (M=19.83, SD=1.50), participated for course credit. All were native speakers of English with normal or corrected-to-normal vision.

Materials. The experimental items included 48 critical items, 6 for each of eight conditions. We adapted the materials from the pupillometry study (Section 4), controlling for word-frequency and length of patient-nouns ('objects') and post-patient contexts (see Table 2): The visual and epistemic object-nouns had very similar mean frequencies

(*M*=84.5, *SD*=70.6 and *M*=82.5, *SD*=69.4, respectively)[13] and mean lengths in terms of number of characters (*M*=8.5, *SD*=2.2 and *M*=6.8, *SD*=2.2, respectively). Neither mean frequencies nor mean lengths differed significantly $t(22)=.07$, $p=.95$ and $t(22)=1.82$, $p=.082$. Similarly, the expressions used in the cancellation region were very similar in terms of mean lengths (visual-consistent *M*=10.7; visual-inconsistent *M*=10.3; epistemic-consistent *M*=9.0; epistemic-inconsistent *M*=10.0) and of the mean frequency of key words (underlined) in each (e.g. post-epistemic: 'ahead of him', 'facing him', 'before him' vs. 'behind him', 'has overcome', 'turned from') (visual-consistent *M*=179.3; visual-inconsistent *M*=202; epistemic-consistent *M*=166; epistemic-inconsistent *M*=158).[14] Critical items employed the same patient nouns as the visual and epistemic items in the pre-study (Section 2.3). There were 48 filler trials.

	Verb	**Object**		**Context**	
Epistemic:					
1. Joe	sees	the problems	that lie	ahead of him.	(s-consistent)
2. Joe	sees	the problems	that lie	behind him.	(s-inconsistent)
3. Joe	is aware of	the problems	that lie	ahead of him.	(s-consistent)
4. Joe	is aware of	the problems	that lie	behind him.	(s-inconsistent)
Visual					
1. Sheryl	sees	the picture	on the wall	behind her.	(s-inconsistent)
2. Sheryl	sees	the picture	on the wall	facing her.	(s-consistent)
3. Sheryl	is aware of	the picture	on the wall	behind her.	(s-inconsistent)
4. Sheryl	is aware of	the picture	on the wall	facing her.	(s-consistent)

Table 2. Example stimuli and regions of interest for eye movement analysis

Apparatus. Eye movements were recorded with an SR Research Ltd. EyeLink 1000 eye-tracker which records the position of the reader's eye every millisecond (see Fn.11). Head movements were minimised with a chin rest. Eye movements were recorded from the right eye. The sentences were presented in 12 pt. Arial black font on a white background.

Design and Procedure. The design was a 2 × 2 × 2 (verb [see/aware] × object [visual/epistemic] × context [s-consistent/s-inconsistent]). All variables were manipulated within subject.

Participants were seated at the eye tracker and instructed verbally. They placed their chins on a chinrest. After a 9-point calibration and validation procedure, participants completed two practice trials and 96 experimental trials. These included 48 critical trials. Each participant saw 'see' and 'aware' versions of critical items in equal number, in each

[13] Here and below, frequency figures refer to occurrence frequencies in the full written and spoken British English reference corpus of Leech et al. (2001).

[14] Since our predictions do not call for comparisons between reading times for verbs, the evident differences in length and frequency between 'see' and 'is aware of' are irrelevant.

condition, as verbs and context-phrases were rotated across lists in a Latin-square design. At the start of each trial, participants were required to fixate a drift-correction dot on the left edge of the monitor, centred vertically. The experimenter then initiated the trial. The sentence appeared after an interval of 500ms and the initial letter of each sentence was displayed in the same position, in terms of x and y coordinates, as the drift correction dot. The entire sentence was presented on a single line on the screen. The participant read the sentence silently and then pressed the spacebar on the keyboard. A plausibility rating prompt appeared, and participants rated sentences' plausibility on a scale from 1 to 5, by pressing the corresponding key on the keyboard. As before, endpoints were explained as 'very implausible' (1) and 'very plausible' (5), and the midpoint (3) as 'neither plausible nor implausible; the decision feels arbitrary'. The entire testing session lasted approximately 30 minutes.

5.3. Results and discussion

For eye movement and plausibility data, we defined outliers as means +/- 3.5 SDs from the mean. There were none. Analyses were conducted with subjects ($F1$) and items ($F2$) as random effects.

Eye movements

Our eye movement findings largely confirmed our predictions, which concerned reading times for object regions [EM1 and EM2] and for post-object context regions [EM3 and EM4]. Our findings were also consistent with the 'good-enough processing' account (Ferreira et al. 2002, Ferreira and Patson 2007), according to which initial shallow processing leads to local interpretations that are subsequently integrated into more comprehensive interpretations, in more in-depth processing (see Section 5.1). Accordingly, we found regressions from the final word of the sentence, in 90% of critical trials, and observed increasing responsiveness to our manipulations, in late than in early eye movement measures. We now first report omnibus tests mandated by the 2×2×2 design, then report how our predictions fared. In a few cases, we examine predictions where the relevant omnibus tests do not provide statistical support for the requisite comparisons.

Object Region

First pass reading times on the object region showed no 3-way interaction and no main effect of verb (*p's*>.25). Crucially, however, the 2 × 2 × 2 (verb × object × context) repeated measures ANOVA did show a main effect of object (see Figure 3): The epistemic objects had longer reading times than the visual objects $F1(1,35)=16.28$, $p<.001$, $\eta^2=.32$, $F2(1,11)=9.55$, $p<.05$, $\eta^2=.47$. As predicted, they did so also specifically in 'see'-sentences (see-visual vs. see-epistemic $t(35)=-2.36$, $p<.05$). Since our norming work excluded frequency and length differences between epistemic and visual objects,

longer reading times for epistemic objects will be due to their more abstract character, which makes them more difficult to process (Binder et al. 2005).

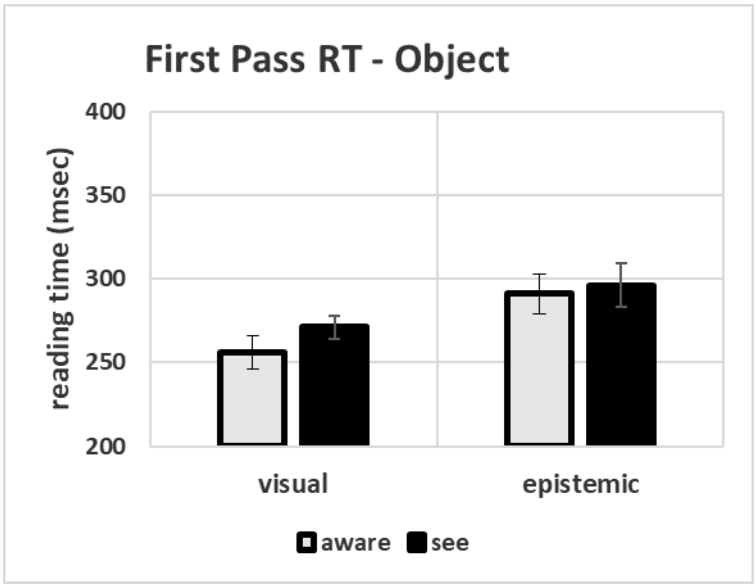

Figure 3. First pass reading times on object region. Error bars show the standard error of the mean.

Second-pass reading times on the object region showed no 3-way interaction (p>.12) but did show a main effect of verb and object, and a by-subjects interaction of object and context $F1(1,35)=5.04$, $p<.05$, $\eta^2=.13$, $F2(1,11)= 2.31$, $p=.16$, $\eta^2=.17$. The object-regions of 'see'-sentences had longer reading times than object-regions in 'aware'-sentences $F1(1,35)= 36.20$, $p<.001$, $\eta^2=.51$, $F2(1,11)= 63.77$, $p<.001$, $\eta^2=.85$. The epistemic objects had longer reading times than the visual objects $F1(1,35)= 44.22$, $p<.001$, $\eta^2=.56$, $F2(1,11)= 12.61$, $p<.01$, $\eta^2=.53$. As predicted by (EM1), this last point also held specifically for 'see'-sentences (irrespective of post-object context) and, yet more specifically, for 'see'-sentences with s-consistent post-object contexts: We found that epistemic objects had longer reading times than visual objects when considering 'see'-sentences irrespective of (i.e. collapsed across) contexts ($t(35)=-4.96$, $p<.001$), and when considering yet more narrowly only 'see'-sentences with s-consistent contexts ($t(35)=7.24$, $p<.001$). As predicted by (EM2), reading times for epistemic objects were longer in 'see'-sentences than in 'aware'-counterparts, irrespective of (i.e. collapsed across) context ($t(35)=-3.97$, $p<.001$). Mean second-pass reading times for epistemic objects were numerically almost identical when followed by s-consistent and s-inconsistent contexts, respectively, in both 'aware'-sentences and 'see'-sentences (Figure 4).

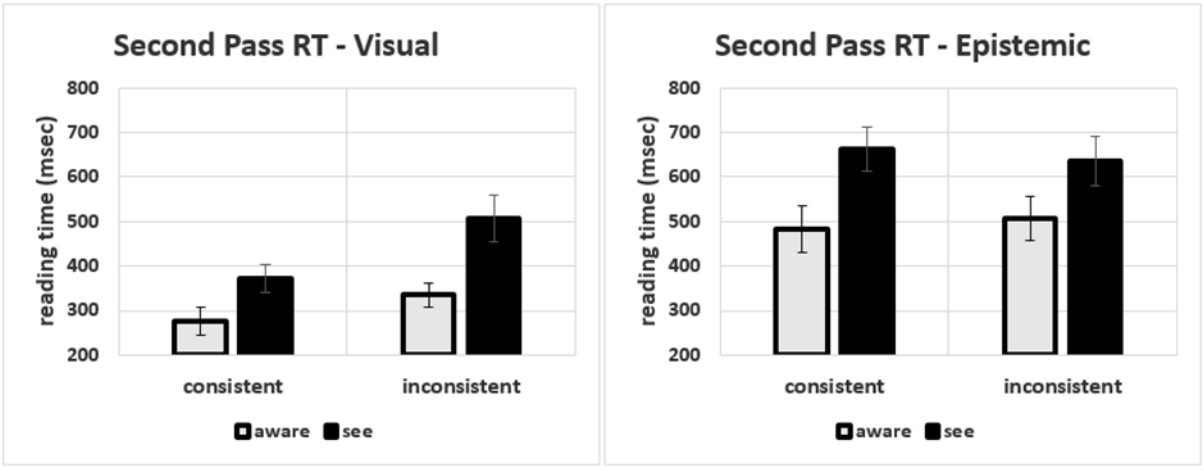

Figure 4. Second pass reading times on object region. Error bars show the standard error of the mean.

Total reading times on the object region showed no 3-way interaction (p>.17) but a main effect of verb and object (see Figure 5). The object-regions of 'see'-sentences had longer reading times than counterparts in 'aware'-sentences $F1(1,35)=35.66$, $p<.001$, $\eta^2=.51$, $F2(1,11)=70.15$, $p<.001$, $\eta^2=.86$,[15] and the epistemic objects had longer reading times than the visual objects $F1(1,35)=54.25$, $p<.001$, $\eta^2=.61$, $F2(1,11)=11.22$, $p<.01$, $\eta^2=.51$. As predicted by (EM1), we also observed longer reading times for epistemic than visual objects when considering specifically 'see'-sentence (irrespective of context) ($t(35)=-5.67$, $p<.001$) and when focussing yet more narrowly on 'see'-sentences with s-consistent contexts ($t(35)=8.25$, $p<.001$). As predicted by (EM2), reading times for epistemic objects were longer in 'see'-sentences than in 'aware'-counterparts, irrespective of (i.e. collapsed across) context ($t(35)=-3.89$, $p<.001$). Mean total reading times for epistemic objects were numerically almost identical when followed by s-consistent and s-inconsistent contexts, respectively, in both 'aware'-sentences and 'see'-sentences (Figure 5).

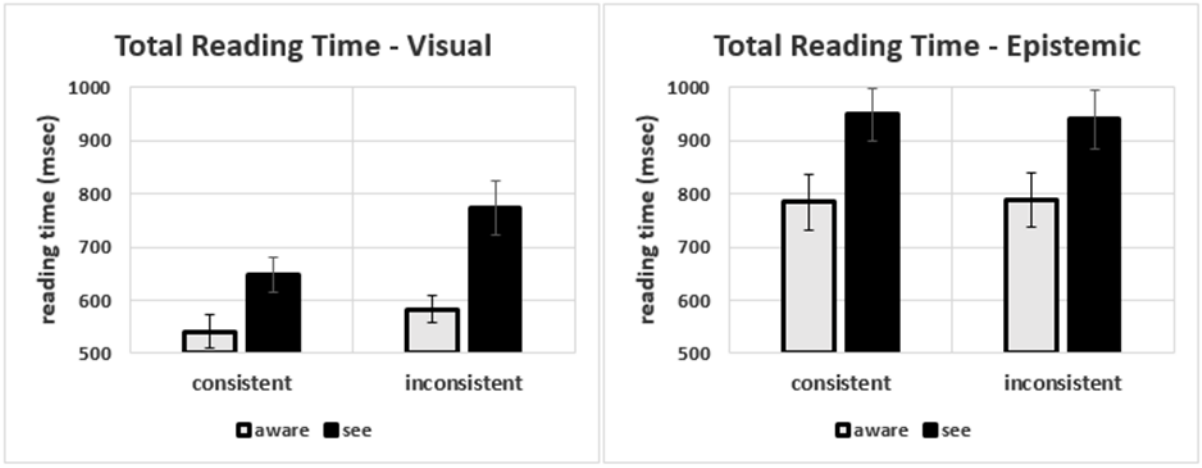

Figure 5. Total reading time on the object region. Error bars show the standard error of the mean.

[15] (EM2) predicts this for sentences with epistemic objects. For a possible explanation concerning sentences with visual objects, see Fn.18.

To sum up, observed reading times for object regions were consistent with the predictions we derived from the hypothesis that the Retention Strategy is employed for interpreting epistemic uses of 'see'. As per prediction (EM1), first-pass, second-pass, and total reading times for the object region were all longer for 'see'-sentences with epistemic than with visual objects. Crucially, this also held specifically for s-consistent sentences, where total object region reading times are not liable to be affected by difficulties to integrate post-object contexts. The fact that total reading times for epistemic objects were the same across sentences with s-consistent and s-inconsistent contexts, in sentences with either verb, further confirms that higher reading times for epistemic than visual objects were not driven by greater difficulties to integrate post-object contexts and increased regressions from such contexts. Finally, as per prediction (EM2), second-pass and total reading times on epistemic objects were longer for 'see' sentences than for their 'aware'-counterparts. These findings support our hypothesis that epistemic uses of 'see' are interpreted with the Retention Strategy (Giora 2003, Giora et al. 2014).

Context region

While our predictions concerning post-object contexts only predict total reading times, we will get a better grasp of both sentence processing and eye tracking measures by considering also first-pass reading times.

First-pass reading times on the post-object context region showed no 3-way interaction ($p>.47$) but a main effect of object and context (see Figure 6). Contexts following visual objects had longer reading times than contexts following epistemic objects (*sic*) $F1(1,34)=11.57$, $p<.01$, $\eta^2=.25$, $F2(1,11)=4.53$, $p=.057$, $\eta^2=.29$, and s-inconsistent contexts had longer reading times than s-consistent contexts $F1(1,34)=15.29$, $p<.001$, $\eta^2=.31$, $F2(1,11)=20.55$, $p<.01$, $\eta^2=.65$. The remaining main effect of verb and the interactions were not significant ($p's>.45$).

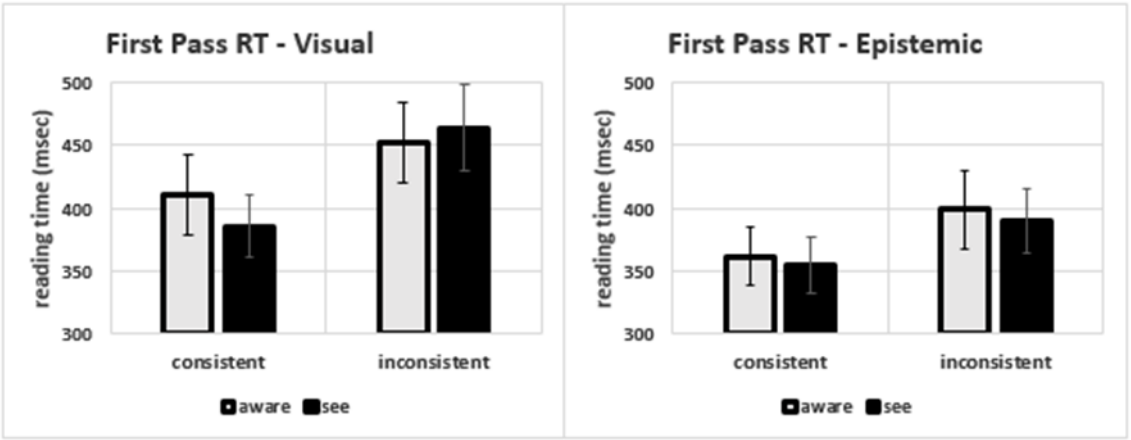

Figure 6. First pass reading times on the context region. Error bars show the standard error of the mean.

While striking *prima facie*, these findings are broadly consistent with the 'good enough processing' account (Ferreira and Patson 2007): At least in first-pass reading, readers only expend effort up to a threshold, tend to process subordinate clauses (like those containing our post-object context regions) superficially (Ferreira and Lowder

2016), and only attempt to integrate information beyond local interpretations where integration is easy enough. Abstract epistemic objects are more difficult to process than concrete visual objects (see above). Our findings suggest that participants made efforts to integrate information from post-object contexts already in first-pass reading only when reading sentences with visual, but not with epistemic objects – where integration efforts get deferred to later processing stages and show up only in second-pass or total reading times. This would account for the observed *longer* reading times for contexts following visual objects.

Total reading times on the context region showed a significant three-way interaction between the variables $F1(1,34)=4.88$, $p<.05$, $\eta^2=.13$, $F2(1,11)=6.64$, $p<.05$, $\eta^2=.38$ as well as a main effect of object and of context (see Figure 7). Context regions following epistemic objects had longer total reading times (significant in the by-subjects analysis) than regions following visual objects $F1(1,34)=5.33$, $p<.05$, $\eta^2=.14$, $F2(1,11)=2.15$, $p=.17$, $\eta^2=.16$, and s-inconsistent contexts had longer reading times than their s-consistent counterparts $F1(1,34)=21.67$, $p<.001$, $\eta^2=.31$, $F2(1,11)=26.92$, $p<.001$, $\eta^2=.71$.

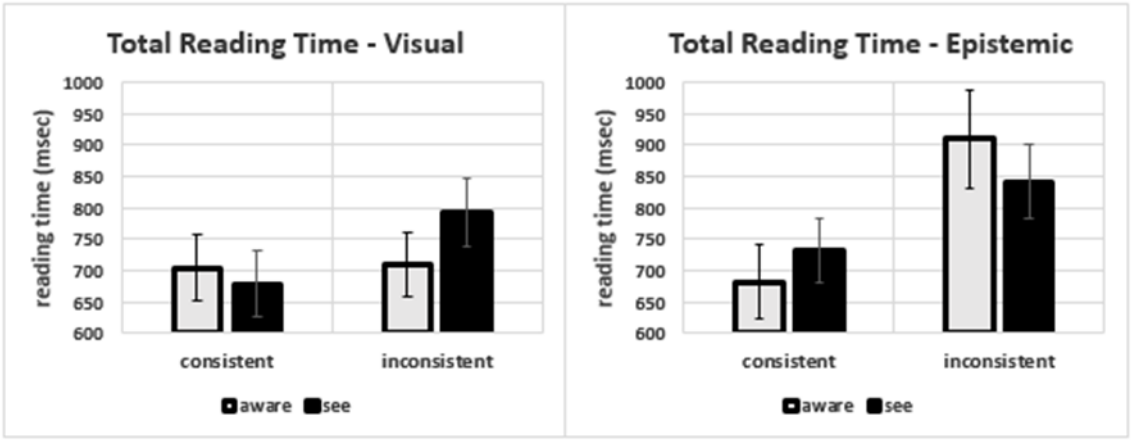

Figure 7. Total reading time on the context region. Error bars show the standard error of the mean.

To decompose the three-way interaction, we considered items with visual and epistemic objects separately. For sentences with visual objects, there was a marginal (by subjects) interaction between context and verb $F1(1,35)=3.32$, $p=.077$, $\eta^2=.09$, though this failed to be confirmed by item analysis $F2(1,11)=1.97$, $p=.19$, $\eta^2=.15$. Neither of the main effects were significant (*p's* >.10). The marginal interaction arose from the fact that total reading times for s-inconsistent contexts in 'see'-sentences were significantly longer than reading times for s-consistent counterparts $t(35)=-2.20$, $p<.05$, and marginally longer than for s-inconsistent contexts in 'aware'-sentences $t(35)=-1.86$, $p=.07$. For sentences with epistemic objects, there was only a main effect of context $F1(1,34)=20.80$, $p<.001$, $\eta^2=.38$, $F2(1,11)=39.17$, $p<.001$, $\eta^2=.78$. Paired comparisons confirmed that s-inconsistent context had significantly longer reading times than s-consistent counterparts, both in 'see'-sentences $t(35)= -2.43$, $p<.05$ and in 'aware'-sentences $t(35)=-4.47$, $p<.001$.

These findings are consistent with our key prediction (EM3) about 'see' sentences: As predicted, total reading times for s-inconsistent contexts were longer than for s-

consistent contexts, in ‘see’-sentences with visual *and* with epistemic objects. Reading times for ‘aware’-sentences, however, did not conform to our expectations: In ‘aware’-sentences with visual objects, total reading times for post-object contexts were not affected by the consistency manipulation, resulting in (marginally) shorter reading times for s-inconsistent contexts than in analogous ‘see’-sentences. By contrast, the consistency manipulation greatly affected context reading times in ‘aware’-sentences with epistemic objects. As a result, total reading times for s-inconsistent contexts were not significantly different for ‘see’- and ‘aware’-sentences with epistemic objects ($t(35)=1.37$, $p=.18$), *pace* (EM4). Indeed, reading times for s-inconsistent contexts in ‘aware’-sentences were even numerically higher than for counterparts in ‘see’-sentences. This finding favours our modified over our initial assumptions about the processing of epistemic uses of ‘is aware of’ (Section 5.1): It suggests that the vision schema is retained also for interpreting such uses of this verb.

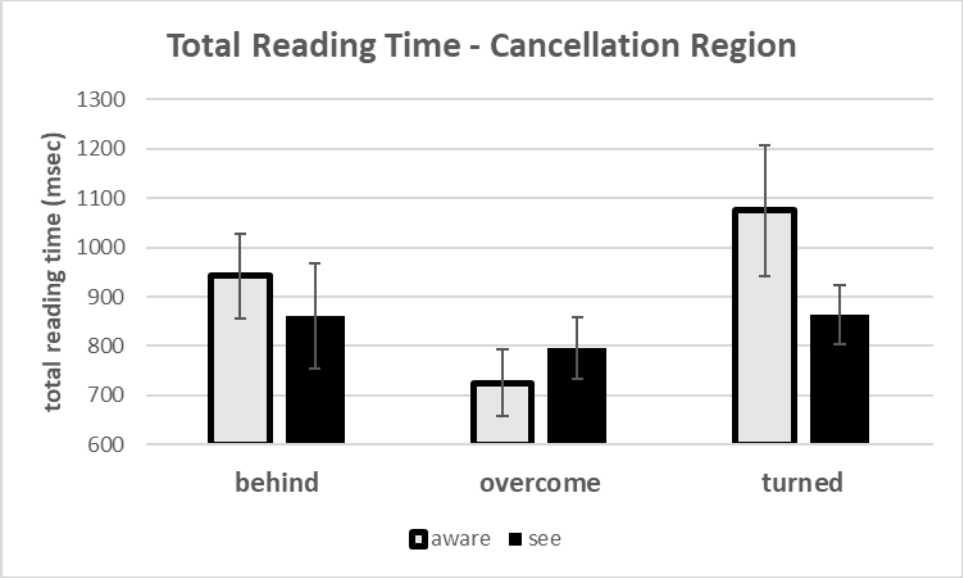

Figure 8. Mean total reading times for the three different cancellation phrases used after epistemic objects. Error bars show the standard error of the mean.

To better understand the processing of post-object contexts, we considered the different cancellation phrases separately. After epistemic objects, we used three different phrases to create s-inconsistent contexts which clash with inferences that *X is in front of S*: ‘behind him’, ‘has overcome’, and ‘turned from’ all place patients behind agents in spatial schemas (*cf.* Section 3.1). Since each phrase was used in just two items, insufficient for a by-items (*F2*) inferential analysis, we only conducted a by-subject (*F1*) analysis. A 2 × 3 (verb × cancellation) repeated measures ANOVA revealed a main effect of cancellation $F(2,70)=4.94$, $p<.05$, $\eta^2=.12$. Contexts with cancellation phrase ‘has overcome’ were read marginally and significantly more quickly than contexts with, respectively, ‘behind’ ($t(35)=1.90$, $p=.066$) and ‘turned from’ ($t(35)= -3.56$, $p<.01$) (Figure 8). Reading times for cancellations ‘behind him’ and ‘turned from’ were not significantly different from each other ($t(35)=-1.07$, $p=.294$). The interaction between variables remained shy of marginal significance $F(2,70)=2.43$, $p=.096$, $\eta^2=.065$,

depriving more detailed comparisons of statistical support. Purely exploratory comparisons between ‘see’- and ‘aware’-sentences suggested a marginal difference for contexts with ‘turned from’ $t(35)=1.83$, $p=.075$.

These differences clearly do not arise from differences in phrase length or word frequency: ‘has overcome’ is the longest phrase, and ‘overcome’ the least frequent word. Higher reading times for ‘turned from’ in ‘aware’- than ‘see’-sentences might arise from activation, by ‘turned from’, of the conceptual metaphor *looking-at as thinking-about* (Fischer 2018). This would lead to a perceived conflict between, e.g. ‘Kelly is aware of the possibilities’ and the implication from ‘she has turned from’, namely, that she no longer thinks about the possibilities. If ‘is aware of’ is more strongly associated with ‘thinks about’ than ‘see’ is, suppression of this previously activated stereotypical associate would lead to higher total reading times on the phrase in ‘aware’- than in ‘see’-sentences. If this is correct, higher reading times in these s-inconsistent ‘aware’-sentences would not be (mainly) due to spatial inferences, and these sentences should be disregarded in assessing our hypotheses. Exclusion of these items yields a mean total reading time of about 800 ms for ‘aware’-sentences with epistemic objects and s-inconsistent contexts. This is numerically below, though not significantly different from, the mean for the corresponding ‘see’-sentences ($t(35)=-.401$, $p=.691$) (Figure 7).

The lower reading times for ‘has overcome’ than ‘lies behind’ may have an explanation in line with our Hypothesis H (Section 2..1): According to H, the salience bias arises where initial activation of contextually inappropriate schema components is not only strong (due to salience) but is also complemented by lateral cross-activation from frequently co-occurring component features of the relevant schema. We now assume that the vision-schema is retained to interpret epistemic uses of both ‘see’ and ‘is aware of’. In instantiations of the vision schema, the spatial-directional feature *X is in front of S* arguably co-occurs frequently with the spatial-vicinity feature *X is near S* (X ‘is around’). Hence these features can be more readily suppressed together than selectively, as one cannot be suppressed completely as long as the other retains activation. Our cancellation phrases all activate spatial schemas serving as source-domain scenarios of conceptual space-time metaphors (Boroditsky and Ramscar 2002, Gentner et al. 2002), but the schemas differ as our phrases carry different literal (source-domain) implications. Whereas ‘X is behind S’ implies that X is still around in the vicinity of S, ‘S has overcome X’ implies X is no longer present to S or around her. The activation of these subtly different schemas therefore either reinforces or inhibits the activation of the component *X is near S* that regularly co-occurs with *X is in front of S*, and thereby hinders or helps suppression of the latter. This would translate into greater integration difficulties and longer reading times for cancellation phrases with ‘behind’ than ‘overcome’.

This explanation can be tested against further data. The account motivates the prediction [Plausibility-3] that s-inconsistent items using the different cancellation phrases should attract different plausibility ratings: By defeating the vicinity-implication, ‘has overcome’ should facilitate complete suppression of the directional feature. S-inconsistent items employing it should therefore be deemed *plausible* (provided s-consistent sentences with epistemic objects are deemed plausible). By reinforcing the

vicinity-implication, ‘behind’ should make complete suppression of the directional feature yet more difficult. S-inconsistent items with it should therefore be deemed distinctly implausible. Finally, ‘S turned from X’ suggests X is still around (though S redirected attention) but implies this less strongly than ‘is behind’ (clearly leaving open the possibility that X moved or vanished since S averted attention). Therefore, ratings of items using ‘turned from’ should attract ratings in between. The fact that the vision schema is more strongly associated with, and hence activated by, ‘see’ than ‘aware’, would predict that suppression effort (evidenced by total reading times) is more successful in ‘aware’-sentences. As a result, ‘aware’-items should be deemed more plausible than ‘see’-counterparts across all cancellation phrases and are likely be placed in a higher plausibility category in the ‘mid-way’ condition with ‘turned from’.

Plausibility

Plausibility results replicated almost perfectly those from the previous pupillometry study (see Section 4) and confirmed our predictions.

A 2 × 2 × 2 (verb × object × context) repeated measures ANOVA showed a significant 3-way interaction $F1(1,35)=22.81$, $p<.001$, $\eta^2=.40$; $F2(1,11)=20.77$, $p=.001$, $\eta^2=.65$, as well as main effects of verb $F1(1,35)=45.71$, $p<.001$, $\eta^2=.57$; $F2(1,11)=117.37$, $p<.001$, $\eta^2=.91$ and context $F1(1,35)=430.87$, $p<.001$, $\eta^2=.93$; $F2(1,11) =104.72$, $p<.001$, $\eta^2=.91$. Sentences with ‘see’ and sentences with s-inconsistent contexts had lower plausibility ratings. To decompose the 3-way interaction, and examine relevant differences, we considered visual and epistemic object-conditions separately (see Figure 9).

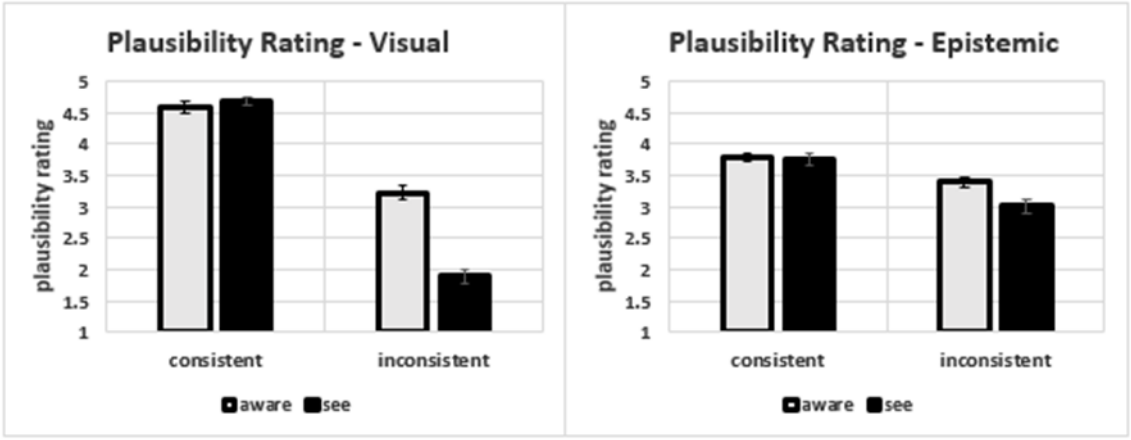

Figure 9. Mean plausibility ratings for each of the eight conditions in the eye tracking study. Error bars show the standard error of the mean.

There were significant 2 × 2 (context × verb) interactions in both the visual object-condition $F1(1,35)=56.89$, $p<.001$, $\eta^2=.62$; $F2(1,11)=190.17$, $p<.001$, $\eta^2=.95$ and the epistemic object-condition $F1(1,35)=14.06$, $p=.001$, $\eta^2=.29$; $F2(1,11)=4.35$, $p=.06$,

η^2=.28.[16] This allowed us to follow up with paired-samples t-tests. As predicted by our first key prediction [Plausibility-1], s-inconsistent 'see'-sentences were deemed less plausible than s-consistent counterparts both when they had visual objects (t(35) =21.87, p<.001) and when they had epistemic objects (t(35)=6.21, p=.001).

Also the comparisons with 'aware'-counterparts turned out as expected: s-consistent 'see'-sentences were deemed equally plausible as 'aware'-counterparts when they took visual objects (t(35)=-.93, p=.36) and when they had epistemic objects (t(35)=.37, p=.72). The context-manipulation then significantly affected the plausibility of 'aware'-sentences only when they took visual objects, and then affected it less than for 'see'-counterparts: 'aware'-sentences with visual objects were deemed less plausible when s-inconsistent than when s-consistent (t(35)=9.66, p<.001). But s-inconsistent 'see'-sentences with visual objects were still deemed less plausible than their 'aware'-counterparts (t(35)=9.16, p<.001). As further expected, the plausibility of 'aware'-sentences with epistemic objects was less strongly affected by the context-manipulation. However, whereas in our previous study (Section 4), s-consistent and s-inconsistent 'aware'-sentences with epistemic objects had attracted numerically almost identical mean ratings, in the present study mean ratings were numerically lower for s-inconsistent sentences than for s-consistent counterparts, and the difference was statistically significant t(35)=3.86, p<.001. Clearly, however, the context-manipulation affected 'aware'- and 'see'-items to a different extent also when they took epistemic objects, and s-inconsistent 'see'-sentences with epistemic objects were rated less plausible than their 'aware'-counterparts (t(35)=4.61, p<.001). Findings thus confirm also our second key prediction [Plausibility-2].

As in the previous study, the predicted plausibility differences translated into categorial differences: Again, s-consistent sentences with either verb and either object were deemed distinctly plausible, that is, attracted plausibility ratings significantly above the neutral mid-point '3' (see-visual: t(35)=27.55, p<.001, aware-visual: t(35)=16.07, p<.001, see-epistemic: t(35)=7.89, p<.001, aware-epistemic: t(35)=10.51, p<.001), as were s-inconsistent 'aware'-sentences with epistemic objects (t(35)=5.24, p<.001). S-inconsistent 'see'-sentences with visual objects were deemed distinctly implausible, with a mean significantly below 3 (t(35)=-10.40, p<.001), while such sentences with epistemic objects were deemed neither plausible nor implausible, with mean ratings not significantly different from '3' (t(35)=.292, p=.772) – as were s-inconsistent 'aware'-sentences with visual objects (t(35)=1.86, p=.072).

To assess our latest prediction [Plausibility-3], we finally considered plausibility ratings for epistemic s-inconsistent items by cancellation phrase (Figure 10). A 2 × 3 (verb × cancellation) repeated measures ANOVA revealed a main effect of verb F(1,35)=21.24, p<.001, η^2=.38, as 'aware'-sentences had higher plausibility ratings than 'see'-counterparts, consistent with previous findings. There was also a main effect of

[16] The marginality of this by-item result is due to the fact that the by-items analysis is less powerful than the by-subjects analysis, involving fewer degrees of freedom. Lower p-values are expected, and marginal by-item results do not impugn the significance of the finding (Cohen 1992).

cancellation $F(2,70)=70.67$, $p<.001$, $\eta^2=.67$, with significant differences between all three cancellation phrases (all *p's* < .05). The interaction between verb and cancellation was not significant ($p > .30$). Finally, all paired comparisons were significant ($p < .05$): In line with prediction [Plausibility-2], s-inconsistent 'see'-sentences of each kind were deemed less plausible than their 'aware'-counterparts. And consistent with prediction [Plausibility-3], items with 'overcome' were rated more plausible than items with 'behind', and 'items with 'turned from' fell between the two.

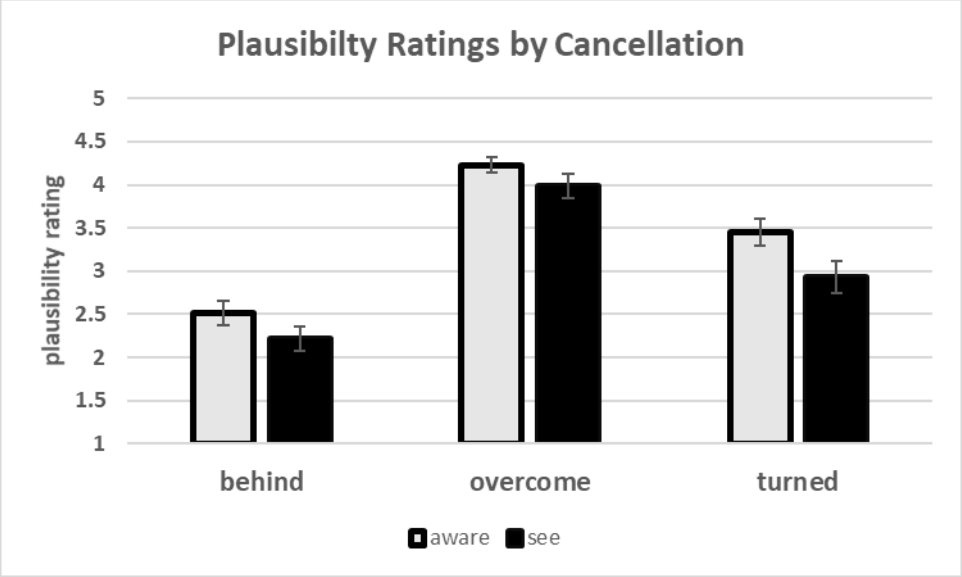

Figure 10. Mean plausibility ratings for see and aware for the three different cancellations. Error bars show the standard error of the mean.

We again conducted one-sample t-tests to determine whether the means were significantly different from neutral mid-point '3'. Prediction [Plausibility-3] was fully borne out: S-inconsistent 'see'- and 'aware'-sentences with cancellation phrase 'behind' were deemed distinctly implausible (significantly below 3: $t(35)=-5.67$, $p<.001$ for 'see', $t(35)=-3.15$, $p<.01$ for 'aware'). By contrast, items with 'overcome' were deemed distinctly plausible (significantly above 3: $t(35)=6.86$, $p<.001$ for 'see', $t(35)=14.29$, $p<.001$ for 'aware'), and attracted mean plausibility ratings that were numerically even higher than those of s-consistent sentences with epistemic objects (*cf.* Figures 9 and 10). Items with 'turned from', finally, were placed into different categories depending upon their verb: 'aware'-items were still deemed distinctly plausible (significantly above 3: $t(35)=3.27$, $p<.01$). But 'see'-items were deemed neither plausible nor implausible (not significantly different from 3: $t(35) =-.39$, $p>.70$).

Discussion

Following on the previous pupillometry study (Section 4), present findings provide further and more detailed evidence that competent speakers make contextually inappropriate stereotypical inferences when the three conditions (i)-(iii) set out by our Hypothesis H are met (see Section 2.1): Our prior work showed that (i) 'see' has a visual sense that is clearly dominant. The confirmation of predictions (EM1) and (EM2)

suggests that (ii) epistemic uses of the verb are interpreted by retaining the situation schema associated with that dominant sense and suppressing contextually irrelevant components of this 'vision schema'. Hypothesis H maintains that where (iii) these irrelevant components continue to receive lateral cross-activation from frequently co-occurring schema components that are contextually relevant, suppression remains partial and contextually irrelevant schema components support contextually inappropriate inferences which influence further cognition (judgment and reasoning). Total reading times in line with prediction (EM3) provide evidence of inappropriate spatial inferences from epistemic uses of 'see'. Plausibility-judgments in line with predictions (Plausibility 1 and 2) provide evidence that inappropriate spatial conclusions influenced further cognition.

Drilling down into differences between specific cancellation phrases provided further support for Hypothesis H, in the shape of evidence for the relevance of condition (iii): Where the cancellation phrase (e.g. 'S has overcome X'), like the previous epistemic object, ruled out as contextually irrelevant *both* of two frequently co-occurring spatial components of the vision schema (*X is in front of S* and *X is near S*), both could be suppressed simultaneously, less suppression effort (reflected in numerically lower total context reading times) led to complete suppression, and no inappropriate inference influenced further cognition – s-inconsistent sentences were deemed as plausible as s-consistent counterparts (as per prediction Plausibility 3). By contrast, where the cancellation phrase failed to rule out one of the two regularly co-occurring spatial features as irrelevant ('turned from') or even implied its relevance ('lies behind'), this feature continued to pass on lateral activation to its regular companion. Accordingly, we observed numerically higher total context reading times and lower plausibility ratings, which provide evidence of inappropriate directional inferences and their influence on subsequent judgment.

The comparison of 'see'- and 'aware'-conditions suggests that salience imbalances (as per condition (i)) are relevant for the cognitive efficacy of the inappropriate inferences examined. Total reading times for object and context regions displayed the same pattern for 'see'- *and* 'aware'-sentences with epistemic objects.[17] This suggests that, contrary to our initial expectations, epistemic uses of 'is aware of' are interpreted, like epistemic uses of 'see', by retaining the vision schema and suppressing its contextually irrelevant component features.[18] The visual use (where the verb takes visual objects) may be the most salient use of 'aware', but is not clearly dominant (see Fn.12), resulting in weaker

[17] For both, s-inconsistency of context increased the reading times for context regions (Figure 7), but not for prior object regions (Figure 5) or verb regions ('see': s-consistent = 350ms, s-inconsistent = 340ms, no difference $t(35)=.34$, $p=.74$; 'aware': s-consistent = 730ms, s-inconsistent = 755ms, no difference $t(35)=-.65$, $p=.52$).

[18] In the visual condition, by contrast, we observed different processing patterns for 'see'- and 'aware'-items. With 'aware', the consistency-manipulation did not affect total reading times for either the object or the context region, but did so for the verb region (s-consistent = 687ms, s-inconsistent = 804ms, a significant difference $t(35)=-2.51$, $p<.05$). With 'see', all three regions had higher total reading times, in s-inconsistent items (verb: s-consistent = 338ms vs. s-inconsistent = 416ms $t(35)= -2.38$, $p<.05$). This suggests that the extra processing effort prompted by s-inconsistency was, in 'aware'-items, devoted to switching to a less salient interpretation of the verb (e.g. from 'is visually aware' to 'has seen and now knows'), but is spread across all three regions in 'see'-items, where no re-interpretation of the verb alone does the trick.

association of the vision schema with ‘aware’ than with ‘see’. Accordingly, similar amounts of suppression effort (as evidenced by increased total reading times for s-inconsistent contexts) led to more complete suppression (evidenced by plausibility ratings), in ‘aware’-sentences: We observed higher plausibility ratings for s-inconsistent epistemic items with ‘aware’ than ‘see’ and, second, found that such ‘aware’-items were deemed distinctly plausible not only when contexts supported suppression through cancellations (‘has overcome’) that explicitly ruled out the contextual relevance of a schema component cross-activating the spatial component. Rather, s-inconsistent ‘aware’-items were also deemed distinctly plausible when the cancellation (‘turned from’) remained silent on the relevance of ‘cross-activators’. Only when the cancellation phrase (‘lies behind’) reinforced the activation of this schema component (*X is near S*), and thereby the lateral cross-activation of the spatial component of interest (*X is in front of S),* did the critical spatial inferences from ‘aware’ go through and affect plausibility judgments. This suggests that, beyond unhelpfully phrased contexts, the Retention Strategy does not make us generally prone to inappropriate inferences which manage to influence further cognition. Rather, our findings provide evidence that it makes us more generally prone to such inferences when the polysemous word at issue displays pronounced salience imbalances and has a dominant sense far more salient than all others.[19]

6. Conclusion

6.1. Main findings

Two studies combining plausibility ratings with pupillometry and eye tracking, respectively, provided evidence of a salience bias in speech and text comprehension: Where a polysemous word has a clearly dominant sense (like ‘see’), utterances that use the word in a less salient sense may trigger contextually inappropriate stereotypical inferences that are licensed only by the dominant sense – and go through to influence further judgment and reasoning, even when hearers/readers know they are inappropriate. Our studies documented inappropriate spatial inferences from epistemic uses of ‘see’ and (to a lesser extent) ‘aware’; in a pre-study, participants drawn from the same population deemed such inferences inappropriate. Our findings suggest that inappropriate stereotypical inferences occur at least where pronounced salience imbalances coincide with an interpretation strategy (‘Retention Strategy’) whereby less salient uses of words are interpreted by retaining the situation schema associated with the most salient sense and attempting to suppress their contextually inappropriate conclusions. Where such attempts remain unsuccessful, e.g. due to continued lateral cross-activation from contextually relevant schema components, competent language users go along with

[19] In line with the adaptive behaviour and cognition programme (Gigerenzer et al. 2011, *cf.* Ferreira and Patson 2007), future research could fruitfully examine to what extent this apparent defect results from a system design that strikes the best balance overall between processing effort and accuracy of information inferred and retained, given real-world task demands.

contextually inappropriate stereotypical inferences, despite knowing they are inappropriate.

6.2. Philosophical relevance

Our findings contribute towards an epistemological profile (Weinberg 2015, 2016) of the key process of stereotypical enrichment. This generally reliable process of automatic inference routinely goes on in language comprehension. It is bound to generate many intuitions thinkers have when considering verbal descriptions of possible cases, in philosophical thought experiments, and to drive many inferences they draw from such descriptions in philosophical arguments. The epistemological profile helps us assess the evidentiary value of these intuitions and to reconstruct such arguments.

Philosophers often take familiar words which ordinary discourse may have endowed with a dominant sense and use them in a special sense (Section 2.2). They may do so to talk about unusual cases which deviate from the stereotype associated with the dominant sense (as envisaged in philosophical thought experiments, for example, about hallucinations or well-behaved zombies). Where this happens, thinkers are liable to make stereotypical inferences which are contextually inappropriate. When the conclusions of such inappropriate inferences strike thinkers as obvious, these intuitions lack evidentiary value. To acquire the right to treat intuitions about unusual (stereotype-divergent) cases as evidence, philosophers need to engage in psycholinguistic investigation at least into the salience structures of the relevant words. Already the first set of jointly vitiating conditions we have identified in the process helps us assess at least some philosophical case-intuitions.

Where inappropriate conclusions are not explicitly endorsed but implicitly presupposed, the finding of the salience bias helps us reconstruct the relevant lines of thought and vindicate reconstructions in the light of plausible principles of charity which permit the attribution of fallacies to competent thinkers only in the presence of empirically supported explanations of why thinkers commit the relevant fallacies under relevant conditions (Thagard and Nisbett 1983). Our specific finding that competent speakers make inappropriate inferences from less salient uses of the verb ‘to see’ helps vindicate our proposed reconstruction of the ‘see-version’ of the ‘argument from hallucination’, which we took to rely on such an inference in its opening step (Section 2.2). Elsewhere (Fischer and Engelhardt, under review) we explain how related salience effects can account for fallacies in versions of the argument that employ the verb ‘is aware of’, instead. These empirically supported reconstructions help resolve this classical paradox and the ‘problem of perception’ (Crane and French 2015, Smith 2002) it engenders together with a parallel paradox (the ‘argument from illusion’, examined in Fischer and Engelhardt, 2016).

6.3. Methodological lessons

Our studies hopefully provide a useful model of how to combine plausibility ratings with pupillometry or reading time measurements, in the cancellation paradigm, to study automatic inferences. In conclusion, we stress three methodological points they may help

illustrate. In the cancellation paradigm, inferences are studied by manipulating the consistency of subsequent text with hypothesised inferences and measuring indices of cognitive effort. Due to 'good enough' processing strategies with initial focus on local interpretations (Ferreira and Patson 2007), increased processing effort engendered by inconsistencies may show up only in later reading time measures (second-pass and total reading times). It may also materialise with delay, namely, on the post-conflict sentence region (Rayner et al. 2004) and at the likely ultimate source of difficulty (on the region regressed to from the conflict region, as with our s-inconsistent visual 'aware'-sentences; see Fn.18).

Second, different eye tracking methods (reading time measurements and pupillometry) can provide complementary evidence but need not yield equivalent results on any specific measure. Our pupillometry study examined increases in mean pupil size between two time-windows, namely, the second half of the sentence and the 1000ms window after sentence offset. Such pupil dilations are indicative of cognitive effort involved in processing the second half of the sentence (Section 4.1). The reading time measure that comes closest to capturing this effort would be summed total reading times for object and context regions (Figure 11).

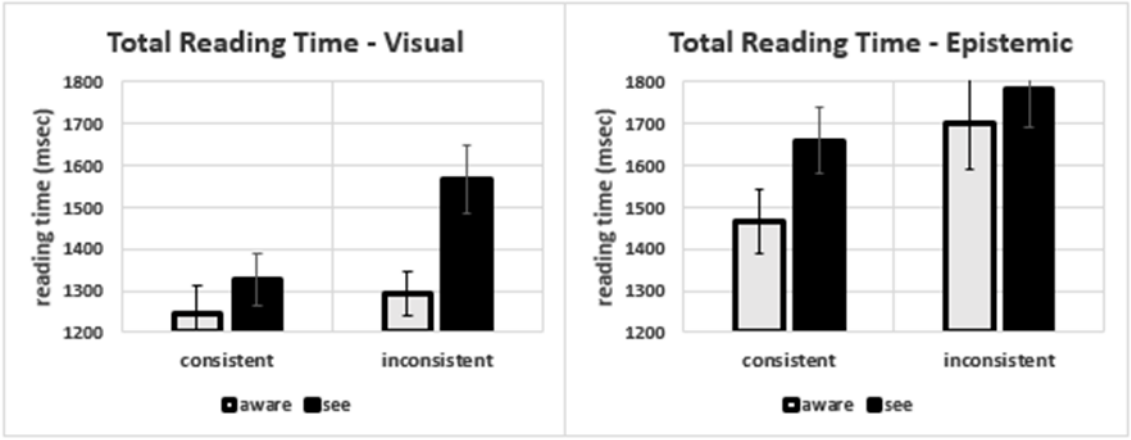

Figure 11. Total reading time on object and cancellation regions. Error bars show standard error of the mean.

We observed pupils dilations for s-inconsistent 'see'- and 'aware'-sentences with visual objects and s-inconsistent 'see'-sentences with epistemic objects. Despite the replication of plausibility results across studies, these dilations are not mirrored in these summed reading times, which are significantly different for visual s-inconsistent 'see'- and 'aware'-sentences, while s-inconsistent 'see'-sentences with epistemic objects do not have significantly longer summed reading times than analogous 'aware'-sentences and s-consistent counterparts.[20] Some differences arise from the facts that in speech

[20] Full analysis of summed total reading times showed a significant 3-way interaction $F(1,35)=7.66$, $p<.01$, $\eta^2=.18$. Follow up 2 × 2 analyses, considering visual and epistemic object conditions separately, showed a significant interaction for visual objects $F(1,35)=5.02$, $p<.05$, $\eta^2=.13$, and for epistemic objects main effects of context $F(1,35)=8.19$, $p<.01$, $\eta^2=.19$ and verb $F(1,35)=4.48$, $p<.05$, $\eta^2=.04$. We observed significant differences between

comprehension (involved in the pupillometry study) there is no 'going back' to sources of difficulty (as in reading) and that pupil dilations and reading times are affected by overlapping but distinct factors (Sections 4.1 and 5.1). More generally, more fine-grained measures need not 'add up' to a global measure and require independent derivation of predictions.

Finally, plausibility ratings and online measures measure different things: In the cancellation paradigm, higher total reading times are indicative of *extent* of suppression and integration effort, at different points. Plausibility ratings reflect *success* of this effort. The two measures hence need not pattern together, since similar effort may lead to more complete suppression of irrelevant schema components, where associations are weaker (as we observed for s-inconsistent epistemic items with 'aware' vs 'see'). Only the plausibility ratings tell us whether an inappropriate inference gets completely suppressed or goes on to influence further cognition. The moment we turn from psycholinguistic questions about sentence processing to experimental philosophy's questions about how automatic inferences affect our judgments and reasoning for better or worse, we need to complement 'online' (process) measures with 'offline' (outcome) measures.[21]

Suggested Readings

Clifton, C., Staub, A., & Rayner, K. (2007). Eye movements in reading words and sentences. In R.P.G. van Gompel et al. (eds.), *Eye Movements. A Window on Mind and Brain* (pp.341–371), Elsevier.

Fischer, E., & Engelhardt, P.E. (2017). Stereotypical inferences: Philosophical relevance and psycholinguistic toolkit. *Ratio*, 30, 411–442.

Laeng, B., Sirois, S., & Gredebäck, G. (2012). Pupillometry: A window to the preconscious? *Perspectives on Psychological Science*, 7, 18–27.

Raney, G.E., Campbell, S.J., & Bovee, J.C. (2014). Using eye movements to evaluate the cognitive processes involved in text comprehension. *Journal of Visualized Experiments*, 83, e50780, doi:10.3791/50780. (with video)

Sirois, S., & Brisson, J. (2014). Pupillometry. *WIREs Cognitive Science*, 5, 679–692.

References

Adler, J.E. (1994). Fallacies and alternative interpretations. *Australasian Journal of Philosophy*, *72*, 271-282.

Allport, D.A. (1985). Distributed memory, modular subsystems and dysphasia. In S. K. Newman and R. Epstein (eds.), *Current Perspectives in Dysphasia* (pp. 207-244). Edinburgh: Churchill Livingstone.

visual-inconsistent see- and aware-items $t(35)=-3.82$, $p<.01$, but not between epistemic-inconsistent see-items and aware-counterparts $t(35)=-.91$, $p>.36$ or epistemic-consistent see-items $t(35)=-1.49$, $p>.14$.

[21] Both authors contributed to material development, study design, and interpretation of results. Paul Engelhardt undertook data-collection and statistical analyses. Eugen Fischer undertook the remaining research. For comments on previous drafts, the authors thank Rachel Giora and Shaun Nichols. For comments on closely related material, we are indebted to audiences in Norwich (July 2017), Osnabrück and Reading (November 2017), and London (June 2018).

Atlas, J. and Levinson, S.C. (1981). It-clefts, informativeness and logical form: radical pragmatics. In P. Cole (ed.), *Radical pragmatics* (pp. 1–62). New York: Academic Press.

Ayer, A.J. (1956/1990). *The Problem of Knowledge*. London: Penguin.

Battig, W.F., & Montague, W.E. (1969). Category norms for verbal items in 56 categories: A replication and extension of the Connecticut category norms. *Journal of Experimental Psychology, 80*, 1-46.

Beatty, J., & Lucero-Wagoner, B. (2000). The pupillary system. In: Cacioppo JT, Tassinary LG, Berntson G, eds. Handbook of Psychophysiology (pp.142-162). Cambridge: CUP.

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.

Bijleveld, E., Custers, R., & Aarts, H. (2009). The unconscious eye opener: Pupil dilation reveals strategic recruitment of resources upon presentation of subliminal reward cues. *Psychological Science*, 20, 1313–1315.

Binder, J.R., Westbury, C.F., McKiernan, K.A., Possing, E.T., & Medler, D.A. (2005). Distinct brain systems for processing concrete and abstract concepts. *Journal of Cognitive Neuroscience, 17,* 905-917.

Boroditsky, L. & Ramscar, M. (2002). The roles of body and mind in abstract thought. *Psychological Science*, 13, 185-188.

Bortfeld, H., & McGlone, M.S. (2001). The continuum of metaphor processing. *Metaphor and Symbol, 16,* 75-86.

Bottini, R. et al. (2015) Space and time in the sighted and blind. *Cognition, 141*, 67–72.

Boyd, K. & Nagel, J. (2014). The reliability of epistemic intuitions. In E. Machery and E. O'Neill (eds.), *Current Controversies in Experimental Philosophy* (pp. 109-127). London: Routledge.

Cappelen, H. (2012). Philosophy without Intuitions. Oxford: OUP.

Carruthers, P. (2002). The cognitive functions of language. *Behavioral and Brain Sciences*, 25, 657–674.

Casasanto, D. & Boroditsky, L. (2008). Time in the mind: Using space to think about time. *Cognition, 106,* 579-593.

Chang, T.M. (1986). Semantic memory: Facts and models. *Psychological Bulletin, 99,* 199-220.

Chisholm, R. (1957). *Perceiving*, Ithaca: Cornell UP.

Chow, W., Smith, C., Lau, E., & Phillips, C. (2016). A 'bag-of-arguments' mechanism for initial verb predictions. *Language, Cognition, and Neuroscience*, 31, 577-596.

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.

Clifton, C., Staub, A., & Rayner, K. (2007). Eye movements in reading words and sentences. In R.P.G. van Gompel et al. (eds.), *Eye Movements. A Window on Mind and Brain* (pp.341–371), Elsevier

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159.

Corbetta, M., Patel, G., & Shulman, G. L. (2008). The reorienting system of the human brain: From environment to theory of mind. *Neuron*, 58, 306–324.

Crane, T., & French, C. (2015). The problem of perception. In N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. Summer 2015. http://plato.stanford.edu/entries/perception-problem/

Deutsch, M. (2015). *The Myth of the Intuitive*. Cambridge, Mass.: MIT Press

Duffy, S., Henderson, J. M., & Morris, R. (1989). The semantic facilitation of word recognition during sentence processing. *Journal of Experimental Psychology: Learning, Memory and Cognition, 15*, 791-801.

Engelhardt, P. E., & Ferreira, F. (2016). Reaching sentence and reference meaning. In P. Knoeferle, P. Pyykkonen, & M. W. Crocker (Eds.), *Visually Situated Language Comprehension.* Amsterdam: John Benjamins

Engelhardt, P.E., Ferreira, F., & Patsenko, E.G. (2010). Pupillometry reveals processing load during spoken language comprehension. *Quarterly Journal of Experimental Psychology, 63,* 639-645.

Evans, J.S.B.T. and Stanovich, K. E. 2013: Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science 8*, 223-241

Farah, M.J., & McClelland, J.L. (1991). A computational model of semantic memory impairment: Modality specificity and emergent category specificity. *Journal of Experimental Psychology: General, 120,* 339-357.

Faust, M., & Gernsbacher, M.A. 1996: Cerebral mechanisms for suppression of inappropriate information during sentence comprehension. *Brain and Language 53*, 234-259

Ferreira, F., Ferraro, V., & Bailey, K. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science*, 11, 11–15.

Ferreira, F., & Lowder, M.W. (2016). Prediction, Information Structure, and Good-Enough Language Processing. *Psychology of Learning and Motivation, 65*, 217-247.

Ferreira, F., & Patson, N. (2007). The 'Good Enough' Approach to Language Comprehension. *Language and Linguistics Compass*, 1, 71–83.

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.

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

Fischer, E. (2018). Two strategies for analogical reasoning: The cases of mind metaphors and introspection. *Connection Science*, 30, 211-243

Fischer, E., & Engelhardt, P.E. (2016). Intuitions' linguistic sources: Stereotypes, intuitions, and illusions. *Mind & Language, 31,* 65-101.

Fischer, E., & Engelhardt, P.E. (2017a). Stereotypical inferences: Philosophical relevance and psycholinguistic toolkit. *Ratio*, 30, 411-442

Fischer, E., & Engelhardt, P.E. (2017b). Diagnostic Experimental Philosophy. *Teorema*, 36 (3), 117-137

Fischer, E., & Engelhardt, P.E. (under review). Lingering stereotypes: Salience bias in philosophical argument.

Fish, W. (2010). *Philosophy of Perception*. London: Routledge.

Frank, S., & Thompson, R. (2012). Early effects of word surprisal on pupil size during reading. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 34, 1554-1559

Frazier, L., and Fodor, J.D. (1978). The sausage machine: a new two-stage parsing model. *Cognition*, 6, 291–325.

Gentner, D., Imai, M., Boroditsky, L. (2002). As time goes by: Evidence for two systems in processing space time metaphors. *Language and Cognitive Processes*, 17, 537-565.

Gerken, M. (2017). *On Folk Epistemology*. Oxford: OUP

Gerken, M., & Beebe, J. (2016). Knowledge in and out of contrast. *Nous*, 50, 133-164.

Gigerenzer, G., Hertwig, R., & Pachur, Th. (2011). *Heuristics : The Foundations of Adaptive Behaviour*. Oxford: OUP

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

Giora, R. & Fein, O. (1999). On understanding familiar and less-familiar figurative language. *Journal of Pragmatics*, 31, 1601-1618.

Giora, R., Fein, O., Aschkenazi, K., and Alkabets-Zlozover, I. (2007a). Negation in context: A functional approach to suppression. *Discourse Processes*, 43, 153–172.

Giora, R., Fein, O., Laadan, D., Wolfson, J., Zeituny, M., Kidron, R., Kaufman, R. & Shaham, R. (2007b). Expecting irony: Context vs. salience-based effects. *Metaphor and Symbol*, 22, 119–146.

Giora, R., Givoni, S. & Fein, O. (2015). Defaultness reigns: The case of sarcasm. *Metaphor and Symbol*, 30, 290-313.

Giora, R., Raphaely, M., Fein, O. & Livnat, E. (2014). Resonating with contextually inappropriate interpretations: The case of irony. *Cognitive Linguistics*, 25, 443-455

Goldberg, A.E. (2003). Constructions: A new theoretical approach to language. *Trends in Cognitive Sciences, 7,* 219–224

Grice, H.P. (1989). Logic and conversation. In his: *Studies in the Ways of Words* (pp. 22-40). Cambridge, Mass.: Harvard UP.

Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. *Science*, 304, 438–441.

Hampton, J.A. & Passanisi, A. (2016). When intensions do not map onto extensions: Individual differences in conceptualization. *Journal of Experimental Psychology: Learning Memory and Cognition*, 42: 505-523.

Hare, M., Jones, M., Thomson, C., Kelly, S., & McRae, K. (2009) Activating event knowledge. *Cognition*, 111, 151-167.

Harley, T.A. (2014). *The Psychology of Language*. 4th ed. London: Psychology Press

Harmon-Vukić, M., Guéraud, S., Lassonde, K.A. & O'Brien, E.J. (2009). The activation and instantiation of instrumental inferences. *Discourse Processes, 46,* 467-490.

Horne, Z., & Livengood, J. (2017). Ordering effects, updating effects, and the spectre of global scepticism. *Synthese*, 194, 1189–1218

Jackson, F. (1977). *Perception. A Representative Theory*. Cambridge: CUP.

Kahneman, D. (1973). *Attention and effort*. Engelwood Cliffs, NJ: Prentice Hall.

Kahneman, D. (2011). *Thinking Fast and Slow*, London: Allen Lane

Kahneman, D. and Frederick, S. 2005: A model of heuristic judgment. In K.J. Holyoak and R. Morrison (eds.), *The Cambridge Handbook of Thinking and Reasoning* (pp. 67-293). Cambridge: CUP.

Kehler, A., Kertz, L., Rohde, H., & Elman, J. L. (2008). Coherence and coreference revisited. *Journal of Semantics, 25,* 1–44/

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.

Kim, A.E: & Osterhout, L. (2005). The independence of combinatory semantic processing: Evidence from anticipatory eye-movements. *Journal of Memory and Language*, 52, 205-225.

Klein, D.E., & Murphy, G.L. (2001). The representation of polysemous words. *Journal of Memory and Language, 45,* 259-282.

Knobe, J. & Nichols, S. (2017). Experimental philosophy. In E.N. Zalte (ed.), *The Stanford Encyclopedia of Philosophy*, Winter 2017. <https://plato.stanford.edu/archives/win2017/entries/experimental-philosophy/>.

Koriat, A. 2007: Metacognition and consciousness. In P. D. Zelazo, M. Moscovitch & E. Thompson (eds.), *The Cambridge Handbook of Consciousness* (pp. 289–326). Cambridge: Cambridge University Press.

Kutas, M. & Federmeier, K.T. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences, 4*, 463-460.

Kutas, M. & Federmeier, K.T. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology, 62,* 621-647.

Laeng, B., Sirois, S., & Gredebäck, G. (2012). Pupillometry: A window to the preconscious? *Perspectives on Psychological Science*, 7, 18–27.

Lakoff, G. (2012). Explaining embodied cognition results. *Topics in Cognitive Science,* 4, 773-785.

Landau, M.J., Meier, B.P. & Keefer, L.A. (2010). A metaphor-enriched social cognition. *Psychological Bulletin*, 136, 1045-1067.

Leech, G., Payson, P. and Wilson, A. 2001: *Word Frequencies in Written and Spoken English: Based on the British National Corpus*. London: Longman

Levinson, S.C. (2000). *Presumptive Meanings. The Theory of Generalized Conversational Implicature*, Cambridge, Mass.: MIT Press.

Levy, B.J., & Anderson, M.C. (2002). Inhibitory processes and the control of memory retrieval. *Trends in Cognitive Sciences, 6,* 299-305.

Lewinski, M. (2012). The paradox of charity. *Informal Logic*, 32, 403-439.

Loewenfeld, I. (1993). *The pupil: Anatomy, physiology, and clinical applications*. Detroit, MI: Wayne State University Press.

Loftus, E.F. (1973). Activation of semantic memory. *The American Journal of Psychology, 86,* 331-337.

Lucas, M. 2000: Semantic priming without association: A meta-analytic review. *Psychonomic Bulletin and Review 7*, 618-630

Machery, E. (2015). The illusion of expertise. In: E. Fischer and J. Collins (eds.): *Experimental Philosophy, Rationalism, and Naturalism* (pp. 188-203). London: Routledge.

Macpherson, F. (2013). The Philosophy and Psychology of Hallucination. In F. Macpherson & D. Platchias (eds.), *Hallucination: Philosophy and Psychology* (pp. 1-38). Cambridge, MA: MIT Press.

Mallon, R. (2016). Experimental philosophy. In H. Cappelen, T. Szabo Gendler, & J. Hawthorne (eds.), Oxford Handbook of Philosophical Methodology (pp. 410-433). Oxford: OUP.

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.

McKoon, G., & Ratcliff, R. (1980). Priming in item recognition: The organization of propositions in memory for text. *Journal of Verbal Learning and Verbal Behavior, 19*, 369-386.

McRae, K., Ferretti, T.R., & Amyote, I. (1997). Thematic roles as verb-specific concepts. *Language and Cognitive Processes, 12*, 137-176.

McRae, K., Hare, M., Elman, J. L., & Ferretti, T. R. (2005). A basis for generating expectancies for verbs from nouns. *Memory & Cognition*, 33, 1174-1184.

McRae, K., & Jones, M. (2013). Semantic memory. In D. Reisberg (ed.), *Oxford Handbook of Cognitive Psychology*, Oxford: OUP.

Mehler, J., Sebastian, N., Altmann, G., Dupoux, E., Christophe, A., & Pallier, C. (1993). Understanding compressed sentences: The role of rhythm and meaning. A*nnals of the New York Academy of Sciences, 682,* 272-282.

Metusalem, R., Kutas, M., Urbach, T.P., Hare, M., McRae, K., Elman, J.L. (2012). Generalized event knowledge activation during online sentence comprehension. *Journal of Memory and Language*, 66, 545-567.

Nado, J. (2014). Philosophical expertise. *Philosophy Compass*, *9,* 631-641.

Nado, J. (2016). Experimental philosophy 2.0. *Thought,* 5, 159–168

Nagel, J. (2012). Intuitions and experiments: a defence of the case method in epistemology. *Philosophy and Phenomenological Research, 85,* 495-527.

Neely, J.H. & Kahan, T.A. (2001). Is semantic activation automatic? A critical re-evaluation. In H.L. Roediger, J.S. Nairne, I. Neath, A.M. Surprenant (eds.), *The Nature of Remembering*. Washington, DC: APA, pp. 69-93.

Nichols, S., & Knobe, J. (2007). Moral responsibility and determinism: The cognitive science of folk intuitions. *Noûs, 41,* 663–685.

Oden, G.C., & Spira, J.L. (1983). Influence of context on the activation and selection of ambiguous word senses. *Quarterly Journal of Experimental Psychology,* 35A, 51–64.

Oppenheimer, D.M. (2006) Consequences of erudite vernacular utilized irrespective of necessity: problems with using long words needlessly. *Applied Cognitive Psychology*, 20, 139–156

Ortony, A. (1993). The role of similarity in similes and metaphors. In A. Ortony (ed.), *Metaphor and Thought*, 2nd edition (pp.342-356). Cambridge: CUP.

Pickering, M.J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences, 36,* 329-347.

Piquado, T., Isaacowitz, D., & Wingfield, A. (2010). Pupillometry as a measure of cognitive effort in younger and older adults. *Psychophysiology*, 47, 560–569.

Pollock, J. (1984). Reliability and justified belief. *Canadian Journal of Philosophy*, 14, 103-114.

Powell, D., Horne, Z., & Pinillos, A. (2014). Semantic integration as a method for investigating concepts. In J. Beebe (ed.), *Advances in Experimental Epistemology* (pp.119-144). London: Bloomsbury.

Raisig, S., Hagendorf, H., & Van der Meer, E. (2012). The role of temporal properties on the detection of temporal violations: insights from pupillometry. *Cognitive Processing*, 13, 83–91.

Rayner, K. 1998: Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin 124*, 372-422

Rayner, K., Kambe, G., & Duffy, S.A. (2000). The effect of clause wrap-up on eye movements during reading. *Quarterly Journal of Experimental Psychology*, *53*, 1061-1080.

Rayner, K., Warren, T., Juhasz, B. J., & Liversedge, S. P. (2004). The effect of plausibility on eye movements in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30, 1290–1301.

Rumelhart. D. E. (1978). Schemata: The building blocks of cognition. In R. Spiro, B. Bruce, & W. Brewer (eds.), *Theoretical Issues in Reading Comprehension*. Hillsdale, N.J.: Erlbaum.

Samuels, E. R., & Szabadi, E. (2008). Functional neuroanatomy of the noradrenergic locus coeruleus: Its role in the regulation of arousal and autonomic function Part I: Principles of functional organisation. *Current Neuropharmacology*, 6, 1–19.

Schwanenflugel, P. J., & Shoben, E. J. (1983). Differential context effects in the comprehension of abstract and concrete verbal materials. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9,* 82-102.

Sereno, S.C., O'Donnell, P.J., & Rayner, K. (2006). Eye movements and lexical ambiguity resolution: Investigating the subordinate bias effect. *Journal of Experimental Psychology: Human Perception and Performance, 32,* 335-350.

Simpson, G.B., & Burgess, C. (1985). Activation and selection processes in the recognition of ambiguous words. *Journal of Experimental Psychology: Human Perception and Performance, 11*, 28-39.

Sirois, S., & Brisson, J. (2014). Pupillometry. *WIREs Cognitive Science*, 5, 679–692

Smith, A.D. (2002). *The Problem of Perception*. Cambridge, Mass: Harvard UP.

Stephens, G.J., Silber, L.J., & Hasson, U. (2010). Speaker-listener neural coupling underlies successful communication. *Proceedings of the National Academy of Sciences, 107*, 14425-14430.

Steenbergen, H. van, & Band, G.P. (2013). Pupil dilation in the Simon task as a marker of conflict processing. *Frontiers of Human Neuroscience*, 7, 215.

Stich, S., & Tobia, K. (2016). Experimental philosophy and the philosophical tradition. In J. Sytsma & W. Buckwalter (eds.), *Blackwell Companion to Experimental Philosophy* (pp.5-21). Wiley Blackwell: Malden.

Swets, B., Desmet, T., Clifton, C., & Ferreira, F. (2008). Underspecification of syntactic ambiguities: Evidence from self-paced reading. *Memory and Cognition, 36*, 201-216.

Tanenhaus, M.K., Carlson, G.N., & Trueswell, J.T. (1989). The role of thematic structures in interpretation and parsing. *Language and Cognitive Processes*, 4, SI 211–234.

Thagard, P. & Nisbett, R.E. (1983). Rationality and charity. *Philosophy of Science*, 50, 250-267.

Thompson, V.A., Prowse Turner, J.A. and Pennycook, G. 2011: Intuition, reason, and metacognition. *Cognitive Psychology 63*, 107-140.

Till, R.E., Mross, E.F., & Kintsch, W. (1988). Time course of priming for associate and inference words in a discourse context. *Journal of Verbal Learning and Verbal Behaviour, 16*, 283-298.

Tulving, E. 2002: Episodic memory: From mind to brain. *Annual Review of Psychology 53*, 1-25

Weinberg, J. (2007). How to challenge intuitions empirically without risking scepticism. *Midwest Studies in Philosophy*, 31, 318-343.

Weinberg. J. (2015). Humans as instruments, on the inevitability of experimental philosophy. In: E. Fischer and J. Collins (eds.), *Experimental Philosophy, Rationalism, and Naturalism* (pp. 171-187). London: Routledge.

Weinberg, J. (2016). Intuitions. In H. Cappelen, T. Szabo Gendler, & J. Hawthorne (eds.), *Oxford Handbook of Philosophical Methodology* (pp. 287-308). Oxford: OUP.

Welke, T., Raisig, S., Nowack, K., Schaadt, G., Hagendorf, H., & van der Meer, E. (2015). Semantic Priming of Progression Features in Events. *Journal of Psycholinguistic Research*, 44, 201–214.

Wheeldon, L.R. & Levelt, W.J.M. (1995). Monitoring the time course of phonological encoding. *Journal of Memory and Language, 34*, 311-334.

Williamson, T. (2007). *The Philosophy of Philosophy*. Oxford: Blackwell.

Zekveld A.A., & Kramer S.E. (2014). Cognitive processing load across a wide range of listening conditions: insights from pupillometry. *Psychophysiology*, 51, 277–284.