Pre-print version of paper accepted (March 4th, 2019) by *Mind and Language*

Lingering stereotypes: Salience bias in philosophical argument

Eugen Fischer and Paul E. Engelhardt

Many philosophical thought experiments and arguments involve unusual cases. We present empirical reasons to doubt the reliability of intuitive judgments and conclusions about such cases. Inferences and intuitions prompted by verbal case descriptions are influenced by routine comprehension processes which invoke stereotypes. We build on psycholinguistic findings to determine conditions under which the stereotype associated with the most salient sense of a word predictably supports inappropriate inferences from descriptions of unusual (stereotype-divergent) cases. We conduct an experiment that combines plausibility ratings with pupillometry to document this 'salience bias': We find that under certain conditions, competent speakers automatically make stereotypical inferences they know to be inappropriate.

1. Introduction

An important strand of experimental philosophy examines whether and when philosophically relevant intuitions have evidentiary value – and whether and when philosophers possess warrant for accepting them (reviews: Mallon 2016, Stich and Tobia 2016). The most ambitious research programme within this remit (initiated by Nichols and Knobe 2007), known as the 'Sources Project' (Pust 2017) or 'cognitive epistemology' (Fischer 2014), develops and experimentally tests psychological explanations of intuitions that help us assess their evidentiary value. A prominent approach is inspired by the heuristics-and-biases program in cognitive psychology (Tversky and Kahneman 1974, Kahneman and Frederick 2005). It seeks to trace specific classes of target intuitions back to automatic cognitive processes which are generally reliable but subject to specific biases.[1] A recent call for an 'experimental philosophy 2.0' (Nado 2016) suggests this approach should be developed further and employed to assess not only intuitive judgments but also philosophical arguments (*cf.* Fischer et al. 2019).

The present paper advances this ambitious agenda by examining a domain-general language process which continually occurs in language comprehension and production, and contributes to shaping inferences and intuitions in many areas of philosophy. Default pragmatic inferences enrich our spontaneous understanding of verbal case descriptions. We examine the most fundamental process of default pragmatic inference, *stereotypical enrichment* (Levinson 2000, Garrett and Harnish 2007). We develop the hypothesis that a cognitive bias affects the generally reliable process of stereotypical enrichment: a *salience bias* that emerges when words with clearly dominant senses are used in a less salient sense, in talk about unusual (stereotype-divergent) cases (Sec.2). A psycholinguistic experiment on inferences from perception-verbs provides evidence for the existence of this bias (Sec.3). We build on the findings from this experiment to analyse philosophical arguments and thought experiments, and develop the *'esotericity thesis'* (Cappelen 2012; Weinberg 2015, Williamson 2007), which holds that automatic inferences and intuitive

[1] For example, intuitive knowledge attributions are traced to a 'mind-reading' competency (Nagel 2012, Gerken 2017) argued to be generally reliable (Boyd and Nagel 2014) but subject to an egocentrism bias (Alexander et al. 2015) and a focal bias (Gerken and Beebe 2016).

judgment about unusual ('esoteric') cases are unreliable (Sec.4). We will see that a far wider range of cases is 'unusual' in the relevant sense than commonly thought, will identify specific conditions under which inferences and judgments about such cases become unreliable, and will suggest ways to avoid the problematic conditions and prevent error.

2. Stereotypical Enrichment and Salience Bias

2.1 Stereotypical Enrichment

Implicit knowledge stored in semantic memory is built up by observation of the co-occurrence of typical properties of things and of typical components of events (McRae and Jones 2013). That implicit knowledge is immediately deployed in language comprehension (Elman 2009, Levinson 2000). Implicit knowledge stored in semantic memory includes stereotypes associated with particular expressions (Hampton 2006). As traditionally conceived, stereotypes represent sets of weighted features which come to mind first, and are easiest to process, when we hear words, and are diagnostic or predictive of the relevant categories. In simple cases, they can be elicited through listing and sentence-completion tasks: 'Tomatoes are___' (e.g., McRae et al. 1997). Priming studies have shown that single words activate stereotypical features rapidly (within 250ms) (review: Engelhardt and Ferreira 2016).

Event nouns (Hare et al. 2009) and verbs (e.g., Ferretti et al. 2001) can be associated with complex stereotypes. When the events or actions denoted typically involve particular kinds of agents, 'patients' acted on, instruments, or relations between them, associated stereotypes include typical features of thematic role-fillers (Tanenhaus et al. 1989). For example, 'frighten' immediately suggests the agent-properties *mean*, *scary*, and *big*, as well as the patient properties *scared*, *small*, and *weak* (McRae et al. 1997), while telic verbs (e.g. 'washing') swiftly activate both initial and resulting patient properties (*dirty*, *clean*) (Welke et al. 2015). These complex stereotypes have internal (thematic) structure and feature activation depends upon thematic fit: Sentence fragments like 'She was arrested by the ___' activate typical agents (*cop*) in post-verbal position only when they leave the agent role blank, but not when they leave open the patient, as in 'She arrested the ___', (Ferretti et al. 2001; *cf.* Kim et al. 2016). Rumelhart (1978) calls these complex, structured stereotypes 'generalized situation schemas' (henceforward 'schemas', for brevity).

Stereotypical associations support spontaneous inferences from words to features stereotypically associated with them. Such spontaneous inferences have been studied through reading time measurements (McKoon and Ratcliff 1980, O'Brien and Albrecht 1992) with eye tracking (Rayner 1998) and ERP studies (Kutas and Federmeier 2011). These studies use a *'cancellation paradigm'*: Participants read sentences where the target expression is followed by a sequel that is inconsistent with (or 'cancels') inferences the participant automatically makes after reading the target expression. Conflicts lead readers to slow down and make more backwards eye-movements, and prompt signature electrophysiological responses (i.e. N400s).[2] For example, when reading 'sewing', people rapidly infer the agent works on cloth and uses a needle – and slow down when the text continues '…the job would be easier if Carol had a needle' (Harmon-Vukič et al. 2009). Verbs thus prompt parallel probabilistic inferences (Ferretti et al. 2001, Welke et al. 2015).

[2] In the cancellation paradigm, N400s are typically taken to indicate violations of readers'/listeners' expectations about the continuation of the sentence. Specifically, N400s result where the continuation violates expectations built on knowledge encoded by stereotypes and schemas (Hagoort et al. 2004), rather than syntactic knowledge.

In addition to schemas associated with individual verbs, we have schemas which encode more general or specific knowledge about recurrent situations (restaurant visits, car inspections, etc.). These are rapidly activated by combinations of verbs and nouns: Participants read the remainder of the sentence more slowly when subject and verb are 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). 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' alone does not prime *paper*) (Matsuki et al. 2011). This suggests that reading activates not only knowledge about typical features of, say, journalists and mechanics, or of checking-events, but also more specific knowledge, e.g., about what mechanics typically check. ERP studies suggest that inferences supported by activation of more specific schemas are made at the earliest possible moment, i.e., right after the verb (Bicknell et al. 2010).

In co-operative communication (Grice 1989), such inferences are made by hearers and anticipated by speakers in line with the neo-Gricean 'I-heuristic' (Levinson 2000; cp. Garrett and Harnish 2007):

(I-speaker) Skip mention of stereotypical features but make deviations from stereotypes explicit.

(I-hearer) In the absence of such explicit indications to the contrary, assume the situation talked about conforms to the relevant schemas, deploy the most specific schemas relevant,[3] and fill in detail in line with this knowledge about situations of the kind at issue.

Explicit marking of stereotype-deviations (as per I-speaker) facilitates suppression of contextually inappropriate inferences (Faust and Gernsbacher 1996). Stereotypical enrichment in line with the I-heuristic is therefore highly context-sensitive. Due to this context-sensitivity, and the relative accuracy of schemas and stereotypes (*cf.* Bordalo et al. 2016) implied by standard accounts of semantic memory (review: McRae and Jones 2013), the process of stereotypical enrichment is generally reliable (*cf.* Fischer and Engelhardt 2016).

2.2 Salience Bias

We now build up to one philosophically relevant set of conditions under which the process of stereotypical enrichment, though generally reliable, leads to inappropriate inferences that affect further judgment and reasoning.

Most words have more than one sense (Klein and Murphy 2001). Whenever we encounter them, the linguistic stimulus activates *all* semantic and stereotypical features associated with the expression, in any of its senses (Simpson and Burgess 1985). It does so regardless of contextual relevance. For example, the word 'mint' activates 'candy', even when used with a less frequent meaning ('All buildings collapsed except the mint') (Till et al. 1988). According to the *Graded Salience Hypothesis* (Fein et al. 2015, Giora 2003), speed and strength of initial activation depend upon the 'salience' of the sense or use, where this label stands for a magnitude unaffected by immediate discourse context. *Salience* (in this sense) is a function of exposure frequency, i.e., of how often the language user encounters the word in this sense. It is further modulated by prototypicality (Rosch 1975), where a sense of a polysemous word (e.g., 'see') is more or less

[3] In incremental utterance interpretation, ever more specific schemas are activated by verbs in conjunction with subject- and object-nouns (see above), with prepositions and syntactic constructions like verb aspect (Ferretti et al. 2007), and with simultaneous visual stimuli (Kamide et al. 2003).

prototypical depending upon whether it stands for more or less prototypical examples of the relevant category (e.g., more or less prototypical cases of *seeing*). The more salient a use is for a hearer, the more rapidly and strongly the situation schema associated with it gets activated.

Let's consider a specific example. The polysemous verb 'to see' exhibits pronounced disparities in salience: Table 1 gives occurrence frequencies in a random sample of 1000 'see'-sentences from the *British National Corpus* as proxy for exposure frequency, and uses production frequencies in a sentence-completion task to measure prototypicality (*cf.* Chang 1986).

Table 1. Occurrence and completion frequencies for 'see' (Fischer and Engelhardt 2017a)

Sense	**Example**	**% of *BNC* occurrences**	**% of completions**
Visual	'I saw him daily.'	68	93.5
Epistemic	'I see your point'	12.4	2.9
Doxastic	'as he saw fit'	9.7	1.9
Phenomenal	'Hallucinating, Macbeth saw a dagger.'	1.1	1.6
Remainder		<5, individually	0

The schema associated with the dominant visual sense (*'vision schema'*) includes agent features such as *S uses her eyes*, *S looks at X*, *S knows X is there*, and *S knows what X is*, and patient features including *X is in front of S*, *X is near S*, *X is there at the same time as S*. Initially, all these features are also strongly activated by less salient senses of "see", such as the epistemic sense ('I see your point').

According to the *Retention/Suppression Hypothesis* (Giora 2003, Giora et al. 2014), the situation schema associated with the word's dominant sense is often deployed to interpret utterances which use the word in a less salient sense: Readers/hearers then retain the dominant schema that is initially strongly activated and suppress its contextually irrelevant components.[4] This pattern has been documented for the interpretation of a variety of figurative uses, including irony (Giora et al. 2007b), sarcasm (Fein et al. 2015), and metaphor (Giora et al. 2007a, Giora & Fein 1999), where it underpins the 'feature transfer' approach of metaphor interpretation (Ortony 1993, Searle 1993). Given the frequency of metaphor alone (Gibbs 1994, 123-4), the Retention/Suppression Strategy is potentially widely used. Applied, for instance, to epistemic uses of 'see' ('Mary sees the possibilities'), the strategy involves retaining the vision schema and trying to suppress all components except *S knows what X is*, to yield the intended interpretation *Mary knows what the possibilities are.* An eye-tracking study with 'see'-sentences provided evidence that this strategy is used to interpret epistemic uses of 'see' (Fischer and Engelhardt, 2019).

When the Retention/Suppression Strategy is applied, and the dominant stereotype (schema) lingers, suppression of some irrelevant component features of this schema will remain partial where a high-frequency word (like 'see') has a dominant sense that is far more salient than the other senses or uses of the word. Because the word is so frequently used, and is used most of the time in that dominant sense, two things will happen: First, the verbal stimulus will strongly activate the features 'central' to the situation schema associated with that dominant use – i.e., those of the schema's component features that are most frequently co-instantiated in the schema's realisations.[5]

---

[4] In contrast, 'direct access accounts' (Gibbs 2002, Vu et al. 2000) hold that a distinct schema is activated when the word is used in a less salient sense.

[5] Not to be confused with the notion of centrality in the concepts literature (e.g., Sloman et al. 1998).

Second, these features will co-occur so often that there will be lateral cross-activation between them, as regularly co-occurring elements of a situation schema activate others (Hare et al. 2009, McRae et. al. 2005). That is, when the word is encountered, all the features central to the schema will be strongly activated and then maintain each other's activation. It will therefore be impossible to completely suppress some of them, while retaining others. Where the Retention/Suppression Strategy requires selective suppression of some, but not all central component features, contextually irrelevant components of a clearly dominant schema will remain partially activated.

Schema components which remain partially activated support inferences that influence further processing. Where contextually irrelevant components of the dominant schema remain partially activated, they support inappropriate inferences that are licensed by the dominant sense of the word but not by the less salient sense that is being employed.

We thus propose the following **Salience Bias Hypothesis (SBH):** When

(i) one sense of a polysemous high-frequency word is much more salient than all others, and
(ii) the dominant situation schema associated with that sense is retained to interpret utterances employing a less salient use of that word, and
(iii) some, but not all, of the central schema components are contextually relevant,

then

(1) contextually inappropriate stereotypical inferences licensed by the dominant sense will be triggered by the less salient use as well, and
(2) these automatic inferences will influence further judgment and reasoning, even when thinkers explicitly know they are inappropriate.

For example, 'Mary sees the possibilities' will not be interpreted simply as 'Mary knows what the possibilities are'; rather, even competent language users will infer, and presuppose in further reasoning, that Mary is looking at the possibilities, that these are located in front of Mary, in her vicinity, etc. But the abstract patient-noun ('possibilities') mostly refers to abstract objects or events lacking spatial location. Where the verb is used with the abstract patient-noun, in the epistemic sense, such spatial inferences will therefore typically lead to conclusions that fail to be true. These inferences are *'contextually inappropriate'*.

Such inappropriate inferences cannot be prevented by explicit marking in the cooperative spirit of the I-heuristic: Speakers can explicitly mark less salient uses of words via riders like 'in a special sense'. Where less salient uses are associated with distinct situation schemas, such marking reinforces the activation of relevant schemas and helps prevent them from being sidelined by dominant schemas, which receive stronger initial activation from the verbal stimulus (Givoni et al. 2013). Such reinforcement can prevent inappropriate inferences. To prevent the inferences posited by the Salience Bias Hypothesis, however, marking would need to reinforce suppression of components of the dominant schema, rather than activation of competitors. We therefore suggest these inferences cannot be prevented by explicit marking and occur even in fully collaborative communication.

The Salience Bias Hypothesis can be examined experimentally by identifying polysemous words and uses that satisfy conditions (i) to (iii), and deriving word-specific hypotheses which can be experimentally tested, e.g., with the cancellation paradigm (Sec. 2.1). Experiments using an outcome measure provided initial evidence for inappropriate spatial inferences from epistemic uses of 'see' (Fischer and Engelhardt 2017a) and for inappropriate doxastic inferences from

phenomenal uses of appearance-verbs (Fischer and Engelhardt 2016) and their robustness in the face of competing pragmatic inferences (Fischer et al. 2019). We now need further evidence from experiments that employ measures which tap into the underlying cognitive process (i.e., also use 'process measures'); we also still need to garner evidence that thinkers know the inferences at issue are inappropriate. We therefore used a novel combination of process and outcome measures (piloted in Fischer and Engelhardt 2017b) in an experiment on inferences from perception-verbs 'see' and 'is aware of', prefaced by a pre-study that examined explicit knowledge.

2.3 Pre-study

Together with our findings about 'see', our general Salience Bias Hypothesis motivates the verb-specific hypothesis that competent speakers will infer spatial patient-properties (*X is in front of S*) also from less salient uses of 'see'. We suggested such inferences occur even where these less salient uses are explicitly marked, and speakers/hearers ought to know the inferences are inappropriate. We therefore conducted a pre-study to obtain premises from which competent speakers know the inferences of interest to be inappropriate, and then tested for such inferences in the main study.

In the pre-study participants assessed spatial inferences from various less salient senses of 'see'. Twenty-three undergraduate psychology students from the University of East Anglia were presented with six spatial inferences each, from perceptual, epistemic, doxastic, and phenomenal uses of 'see', as well as from visual and epistemic uses of 'aware' (see Table 2 for examples).

Table 2. Inferences assessed in pre-study.

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

Participants were asked to indicate on a 5-point scale 'how confident you are that in situations where the first sentence [premise] is true also the second sentence [conclusion] will typically be true'. Mid- and endpoints were explained as follows:

> '5 means you are very confident that the second sentence will typically be true' in those situations.
>
> '1 means that you are very confident that the second sentence will typically fail to be true in [those] situations […] This may be the case because the second sentence will typically be false in such situations or because it makes no sense.'
>
> '3 means you do not feel confident either way: you can't tell whether the second sentence will be true or not, in situations where the first sentences is true.'

Participants were instructed to interpret each premise 'so that it makes good sense to you', and to always interpret conclusions literally. The pen-and-paper questionnaire included two practice items and six critical items for each use of interest.

Mean ratings for spatial inferences from visual uses of 'see' (4.7) and 'aware' (3.69) were significantly above neutral mid-point '3' ($t(22)=27.49, p<.001$ and $t(22)=4.08, p<.001$, respectively): Participants were confident these inferences would typically lead to true conclusions. By contrast, spatial inferences from phenomenal (2.19) and doxastic (1.71) uses of 'see' were rated significantly below mid-point ($t(22)=-5.07, p<.001$ and $t(22)=-13.20, p<.001$, respectively), and inferences from

epistemic uses of 'see' (1.58) and 'aware' (1.53) were rated yet lower ($t(22)=-11.46$, $p<.001$ and $t(22)=-10.47$, $p<.001$, respectively; 64% of responses to epistemic uses of 'see' were '1', as were 66**%** of responses to such uses of 'aware'): Participants were very confident that spatial inferences from these non-visual uses typically lead to conclusions that fail to be true.

Pre-study participants were most confident that spatial inferences from epistemic uses are contextually inappropriate in this way. The epistemic sense of 'see' is also the most salient of the verb's less salient senses (Table 1). Positive findings of inappropriate automatic inferences from this sense are therefore most likely to carry over to the others, and our main study used participants from the same population to examine the following hypotheses:

H_1 Competent speakers automatically 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 influence subsequent judgment and reasoning, regardless of contextual (im)propriety.

3. Experiment

To examine whether inferences are made and influence further cognitive processing, we combined a plausibility rating task with pupillometry, in a cancellation paradigm. In this paradigm, readers or listeners must expend cognitive effort to deal with conflicts between automatic inferences and textual sequels that cancel them (see Sec. 2.1). Pupil dilation is an index of cognitive effort (Kahneman 1973). These pupil responses are spontaneous and impossible to suppress at will (Loewenfeld 1993), 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 offer a window into preconscious automatic processing (Laeng et al. 2012, Sirois and Brisson 2014) which is inaccessible to pure outcome measures (like plausibility ratings).

3.1 Predictions

Our participants heard and then rated sentences like the following on a 1-5 plausibility 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'*)

The final part of the sentence is either consistent or inconsistent with the 'see'-stereotype (vision schema) and, specifically, with the hypothesised spatial inferences from 'S sees X' to *X is in front of S*. If these inferences are made, their conclusions' clash with s-inconsistent sequels will engender comprehension difficulties that cost cognitive effort to overcome and trigger pupil dilations. The pupil takes approximately 1 second to expand to maximum size following a point of difficulty (Engelhardt et al. 2010), so clashes will trigger dilations during the 1 second *'offset time window'* after participants heard the sentence. Significant dilations for s-inconsistent sentences absent for s-consistent sentences would provide evidence of the hypothesised automatic inferences.

To show that initial inferences are supported by (stereotypical) features of the verb, we manipulated not only the post-verbal context but also the verb. As contrast, we used '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*).[6] In a prior production experiment with a

[6] https://www.macmillandictionary.com/dictionary/british/aware, http://wordnetweb.princeton.edu/perl/webwn, last accessed February 22, 2019.

sentence-completion task (see Sec. 2.2), 'aware of' was paired less than half the time (46%) with visual objects as patients/themes, which agents would be aware of in virtue of looking at them, while it was so paired only 23% of the time in a random sample of 1000 'aware of' sentences from the *British National Corpus*. We inferred that the vision schema and, specifically, spatial patient-properties would be associated less strongly with 'aware' than with 'see'. We therefore also examined sentences 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.

and predicted participants would respond differently to s-inconsistent 'see'-sentences than to their 'aware'-counterparts (like 2b, which we now also refer to as 's-inconsistent', viz. inconsistent with the 'see'-stereotype) (below).

According to our hypotheses, epistemic uses of 'see' should also prompt spatial inferences, whose conclusions influence further reasoning. In the absence of contextual cues, concrete patient-nouns ('picture', 'car') invite visual interpretations of 'see', and abstract patient-nouns ('challenges', 'opportunities', henceforth 'epistemic objects', for convenience), whose referents typically cannot be literally 'seen', but known, invite epistemic readings. We therefore also manipulated the object, by including 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'). Our pre-study participants regarded spatial inferences from epistemic items as inappropriate (Sect. 2.3). If these inferences are automatically made anyway, they will lead to comprehension difficulties in s-inconsistent items absent from s-consistent counterparts: Though conventional, the present space-time 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 our metaphorical sequels will initially activate spatial schemas that place objects in front of, or behind a forward/future-facing subject (Gentner et al. 2002). If inappropriate spatial inferences from the prior verb are made, their conclusions will therefore engage spatial reasoning, will clash with s-inconsistent sequels ('s/he left behind') and will engender comprehension difficulties indicated by pupil dilations.[7]

Our hypothesis H_1 about stereotypical inferences from 'see' therefore predicts that

[*Prediction Pupil*] s-inconsistent 'see'-sentences with visual *and* with epistemic objects will prompt pupil dilations following sentence offset.

Since pairings of verbs and patient-nouns swiftly activate more specific generalized situation schemas (Sec. 2.1), we assume that 'is aware of', followed by a visual object, will activate the vision-

[7] Section 3.4 will exclude alternative explanations suggesting apparent clashes are generated without spatial inference. Norming work excluded spatial orientation of patient-nouns as confound.

schema (agent is aware of patient because she sees it), and further expect that also s-inconsistent 'aware'-sentences with visual objects will prompt spatial inferences and pupil dilations in the offset window. With epistemic objects, by contrast, we expect 'aware' will activate only a purely epistemic sense. Accordingly, we do not expect s-inconsistent 'aware'-sentences with epistemic objects to prompt spatial inferences and dilations. If predicted dilations are indeed due to spatial inferences from the main verb (or verb-cum-object-noun) s-consistent sentences with either verb will not prompt dilations.

Automatic inferences in language comprehension (as picked up by pupillometry) need not influence further judgment and reasoning: Where they conflict with discourse or world knowledge that gets swiftly activated in language comprehension (Metusalem et al. 2012), conclusions may get completely suppressed from the situation model constructed by the comprehender, within one second of sentence offset (Fischer and Engelhardt 2017b, Exp.2; *cf.* Faust and Gernsbacher 1996). To examine our hypothesis H_2 that contextually appropriate *and* inappropriate inferences influence further judgment, we therefore combined pupillometry with plausibility ratings which were prompted one second after sentence offset: If spatial conclusions are not completely suppressed, the situation models constructed for s-inconsistent items will remain inconsistent, resulting in lower plausibility ratings for s-inconsistent than s-consistent items.[8]

To spell this out for the crucial epistemic items: If contextually inappropriate inferences from epistemic uses get made (as per H_1) but their conclusions are completely suppressed from the situation model, participants will base their plausibility ratings on purely metaphorical interpretations of items. A norming study confirmed that, on such an interpretation, s-*in*consistent epistemic items (*Jack knows what problems he had in the past*) are deemed *more* plausible than s-consistent counterparts (*Joe knows what problems he will have in the future*) (see Appendix A). By contrast, if (as per H_2) spatial conclusions inferred from epistemic uses of 'see' are only partially suppressed, they will remain part of the situation model and prevent purely metaphorical interpretation of the item. The resulting more literal utterance interpretations will be more plausible for s-consistent items (*Mountaineer Joe visually discerns the problems – crevices, etc. – that lie ahead on his route*) than their s-inconsistent counterparts. H_2 thus predicts that

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

With *visual* objects, 'aware'-sentences will activate the vision schema and prompt spatial inferences (above). However, since this schema is less strongly associated with 'aware', the features rendered irrelevant by s-inconsistent contexts (e.g., *S looks at X*) are more completely suppressed, securing a more clearly epistemic interpretation of s-inconsistent items (*Chuck knows there is a spot on the wall behind him*), on which they are more plausible than 'see'-counterparts (perhaps Chuck knows because he looked at the wall before, or was told). Since, with *epistemic* objects, we expect 'aware' will prompt no spatial inferences (above), we expect the context-manipulation not to affect plausibility ratings of epistemic 'aware'-items. By contrast, H_2 predicts spatial conclusions from epistemic uses of 'see' will linger, despite their impropriety, and influence plausibility ratings. Where s-consistent 'see'-sentences are as plausible as 'aware'-counterparts, the context-manipulation will hence make s-inconsistent 'see'-sentences less plausible than aware-counterparts.

[8] Fischer & Engelhardt (2019) explain in detail how to interpret plausibility ratings on content- and experience-based accounts of metacognitive judgment, and in light of potential confounds.

H_2 thus predicts that

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

3.2 Methods

4.2.1 Participants. As in the pre-study, undergraduate psychology students from the University of East Anglia participated for course credit. All thirty-eight were native speakers of English.

4.2.2 Materials. The experimental items included 48 critical items, viz., 6 for each of the eight conditions (examples 1a – 4b). Appendix B provides a list of critical items and further details.

4.2.3 Design and Procedure. In a 2×2×2 design, we manipulated context (s-consistent/s-inconsistent), verb (see/aware), and object (visual/epistemic), within-subject.

Participants heard items while looking at a fixation cross on a computer screen and rated them from '1' ('very implausible') to '5' ('very plausible'), once the fixation cross was replaced by a rating prompt. Mean pupil diameter was measured with an Eyelink 1000 during the second half of the sentence and during the second after sentence offset. Analyses were conducted with subjects (*F*1) and items (*F*2) as random effects. Appendix C provides further details.

3.3 Results

3.3.1 Plausibility: A 2×2×2 (context × verb × object) repeated measures ANOVA showed a significant 3-way interaction $F1(1,37)=17.49$, $p<.001$, $\eta^2=.32$; $F2(1,11)=11.23$, $p=.006$, $\eta^2=.51$, as well as main effects of verb $F1(1,37)=45.05$, $p<.001$, $\eta^2=.55$; $F2(1,11)=51.33$, $p<.001$, $\eta^2=.82$ and context $F1(1,37)=147.18$, $p<.001$, $\eta^2=.80$; $F2(1,11)=224.96$, $p<.001$, $\eta^2=.95$. 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 1). Where the 2×2 (context × verb) interaction was significant, we followed up with paired-samples *t*-tests. There were significant interactions with both visual objects $F1(1,37)=62.43$, $p<.001$, $\eta^2=.63$; $F2(1,11)=47.64$, $p<.001$, $\eta^2=.81$ and epistemic objects $F1(1,37)=5.27$, $p=.027$, $\eta^2=.13$; $F2(1,11)=7.23$, $p=.021$, $\eta^2=.40$. With visual objects, s-inconsistent 'see'-sentences were deemed less plausible than s-consistent counterparts $t1(37)=18.63$, $p<.001$, $\eta^2=.90$; $t2(11)=23.13$, $p<.001$, $\eta^2=.98$, as per prediction [Plausibility-1]. Also s-inconsistent 'aware' sentences were deemed less plausible than s-consistent counterparts $t1(37)=7.29$, $p<.001$, $\eta^2=.59$; $t2(11)=7.56$, $p<.001$, $\eta^2=.84$, but the difference was less pronounced. While s-consistent sentences with 'see' and 'aware' were deemed equally plausible $t1(37)=-1.59$, $p=.12$, $\eta^2=.06$; $t2(11)=-1.32$, $p=.21$, $\eta^2=.14$, s-inconsistent 'see' sentences were rated less plausible than 'aware'-counterparts $t1(37)=9.08$, $p<.001$, $\eta^2=.69$; $t2(11)=8.02$, $p<.001$, $\eta^2=.85$, as per prediction [Plausibility-2].

This pattern was reproduced by sentences with epistemic objects, to the extent predicted by H_2. Again, s-inconsistent 'see'-sentences were deemed less plausible than s-consistent counterparts $t1(37)=4.29$, $p<.001$, $\eta^2=.33$; $t2(11)=2.87$, $p=.015$, $\eta^2=.43$, as predicted [Plausibility-1]. By contrast, the plausibility of 'aware'-sentences with epistemic objects was not affected by context: As expected, s-consistent and s-inconsistent 'aware'-sentences with epistemic objects were regarded as equally plausible $t1(37)=.15$, $p=.88$, $\eta^2=.00$; $t2(11)=-.096$, $p=.925$, $\eta^2=.00$. Again, s-consistent sentences with 'see' and 'aware' were deemed equally plausible $t1(37)=.39$, $p=.70$, $\eta^2=.00$; $t2(11)=.188$, $p=.854$, $\eta^2=.00$, while the s-inconsistent 'see'-sentences were rated less plausible than 'aware'-counterparts $t1(37)=3.12$, $p<.01$, $\eta^2=.21$; $t2(11)=7.95$, $p<.001$, $\eta^2=.85$, as

predicted [Plausibility-2].

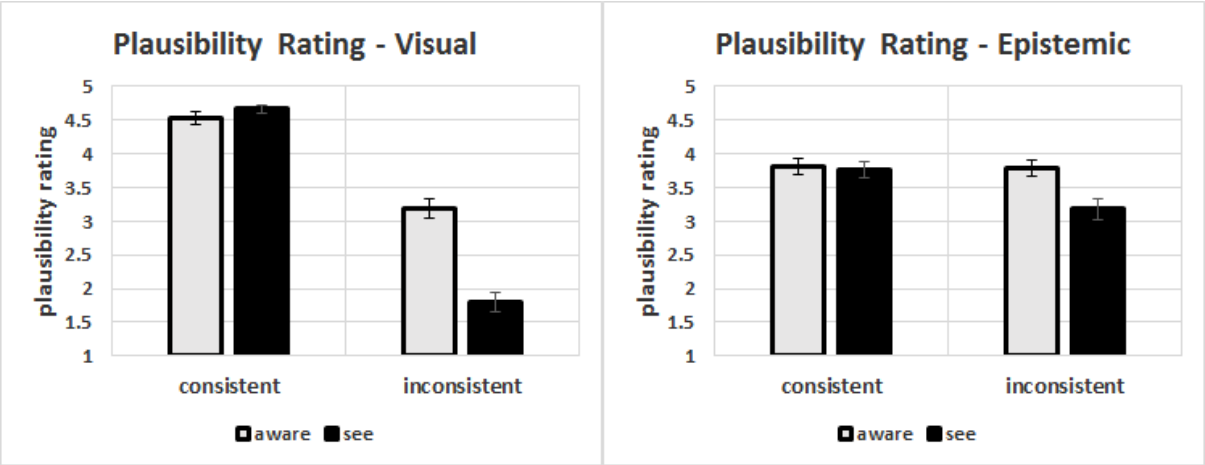

Figure 1. Mean plausibility ratings for each of the eight conditions. Error bars show the standard error of the mean.

In all four s-consistent conditions, sentences were judged plausible, with mean ratings significantly above the neutral mid-point '3'. This was assessed with one-sample *t*-tests with a test value of 3 (aware-visual: $t1(37)=16.74$, $p<.001$, $t2(11)=12.63$, $p<.001$; aware-epistemic: $t1(37)=6.29$, $p<.001$, $t2(11)=6.00$, $p<.001$; see-visual: $t1(37)=27.49$, $p<.001$, $t2(11)=23.59$, $p<.001$; see-epistemic: $t1(37)=6.58$, $p<.001$, $t2(11)=9.54$, $p<.001$). The same occurred for s-inconsistent 'aware'-sentences with epistemic objects $t1(37)=6.77$, $p<.001$, $t2(11)=5.51$, $p<.001$. By contrast, participants rated s-inconsistent 'aware'-sentences with visual objects and s-inconsistent 'see'-sentences with epistemic objects as neither plausible nor implausible: Their mean ratings were not significantly different from 3 (aware: $t1(37)=1.31$, $p=.20$, $t2(11)=.988$, $p=.35$; see: $t1(37)=1.19$, $p=24$, $t2(11)=.350$, $p=.35$). Finally, s-inconsistent 'see'-sentences with visual objects were deemed distinctly implausible, with mean ratings significantly below 3 $t1(37)=-8.63$, $p<.001$, $t2(11)=-9.80$, $p<.001$. Participants' ratings thus placed items into three distinct categories (plausible, neutral, and implausible), and all predicted plausibility differences translated into categorical differences.

3.3.2 Pupillometry: We assessed pupil diameter during the 1000 ms time window after sentence offset. A 2×2×2 (object × context × verb) repeated measures ANOVA showed a significant main effect of context $F1(1,37)=11.26$, $p<.01$, $\eta^2=.23$; $F2(1,11)=7.98$, $p=.017$, $\eta^2=.42$. The s-inconsistent items resulted in larger pupil diameters (see Figure 2). There was also a significant (by subjects) interaction between object and context $F1(1,37)=4.67$, $p=.04$, $\eta^2=.11$; $F2(1,11)=4.63$, $p=.054$, $\eta^2=.30$.[9] However, the 3-way interaction was not significant (*p's*>.80). The interaction between object and context was primarily driven by differences in sentences with visual objects and in sentences with s-inconsistent contexts (see Figure 3): Considering sentences with visual objects, paired samples *t*-tests revealed significant differences between sentences with s-consistent and s-inconsistent contexts $t1(37)=-3.92$, $p<.001$, $\eta^2=.29$; $t2(11)=-4.17$, $p=.002$, $\eta^2=.61$. Considering s-inconsistent sentences, we found significant by-subject differences between sentences with epistemic and with visual objects $t1(37)=-2.12$, $p=.04$, $\eta^2=.11$, but this was not confirmed by item analysis $t2(11)=.579$, $p=.574$, $\eta^2=.03$. The other two paired comparisons were not significant

[9] The marginality of the by-item result and one below (both .054), results from the lower number of relevant items which renders the by-items analysis less powerful than the by-subjects analysis (lower degrees of freedom). Lower *p*-values are expected, and marginal by-item results do not impugn the finding's significance (Cohen 1992).

(*p*'s>.05).

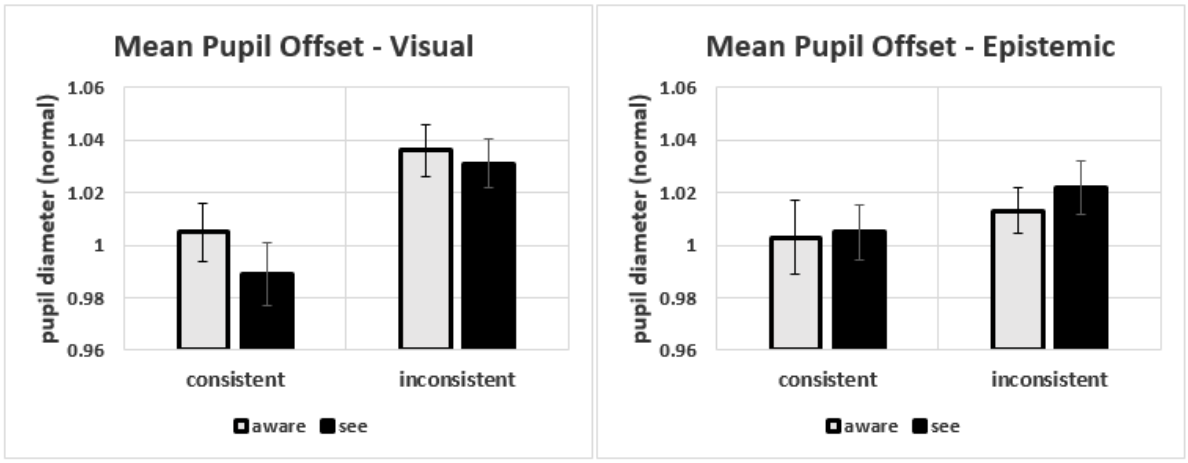

Figure 2. Baseline adjusted pupil diameter in 1000 ms time window following sentence offset. Error bars show the standard error of the mean.

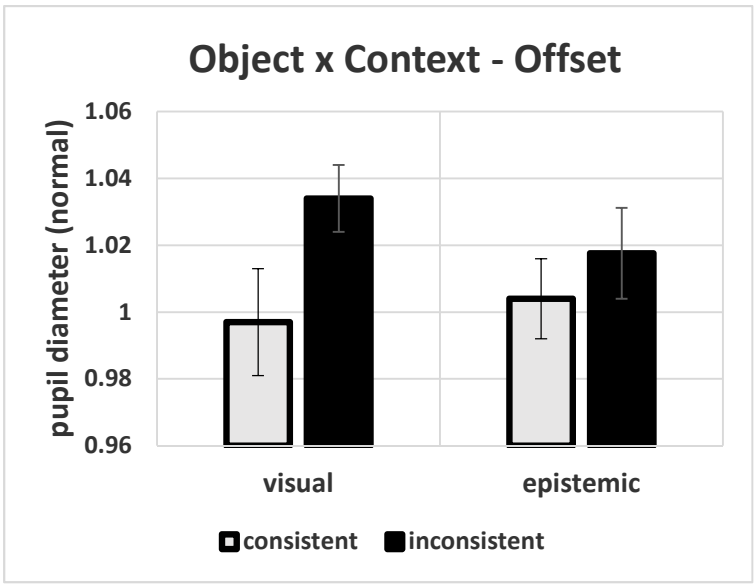

Figure 3. Baseline adjusted pupil diameter showing the object by context interaction. Error bars show the standard error of the mean.

Crucially, follow-up analyses considered whether the pupil diameter was significantly larger in the offset time window compared to the previous time window. To this end, we conducted one-sample *t*-tests with a test value of 1. A value of 1 would indicate that mean pupil diameter was the same during offset as during the second half of the sentence. As predicted [Prediction Pupil], participants' pupil size significantly increased after hearing s-inconsistent 'see'-sentences with visual objects $t1(37)=3.27, p<.01, t2(11)=3.28, p<.01$ *and* with epistemic objects $t1(37)=2.17, p=.037$, $t2(11)=3.41, p<.01$. As expected, there was also increased pupil size after hearing s-inconsistent 'aware' sentences with visual objects $t1(37)=3.72, p<.01, t2(11)=2.16, p=.054$. As further expected, none of the other one-sample *t*-tests were significant (all *p*'s>.14).

3.4 Discussion

Pupil results for 'see'-sentences were consistent with predictions, as were all plausibility ratings. Pupil results provide evidence that spatial inferences were made from visual *and* epistemic uses of

'see' (as per our H_1); plausibility results suggest these inferences influenced further cognition, in both cases (as per H_2). By contrast, pupil results for 'aware'-sentences slightly diverged from expectations: We did not expect increased pupil size for s-inconsistent 'aware'-sentences with epistemic objects. Here, pupil size indeed was not significantly greater than 1, but nonetheless fell clearly between the other conditions with epistemic objects. This data point substantially affected the significance of the 3-way interaction and left us without statistical support for paired comparisons between dilations for 'see' and 'aware'.

One possible explanation for this slight (numeric) increase in pupil size is that 'aware' is weakly associated with the vision schema, and thus weakly activates this schema, even without a visual patient-noun. This would be consistent with the fact that, in sentence-completion tasks, participants completed 'aware'-stems with perceptual objects, on average, almost half the time (Fischer and Engelhardt 2017a, Exp.1). Given weaker activation of the vision schema by 'aware' than 'see', less suppression effort results in complete suppression of contextually irrelevant components of the vision schema, evidenced, respectively, by dilations shy of significance and high plausibility ratings for s-inconsistent 'aware'-sentences with epistemic objects.

Plausibility ratings mirrored pupillometry results in that precisely in the three conditions with significant pupil dilations participants refrained from rating sentences as 'plausible'. The absence of more detailed mirroring coheres with our view that pupillometry garners evidence of initial automatic inferences, whereas plausibility ratings are based on subsequently built situation models whose construction takes into account world knowledge and background beliefs, and may involve different degrees of suppression of initial conclusions (Fischer and Engelhardt 2017b).

Where our verbs take visual objects, spatial conclusions (*X is before S*) will get suppressed in interpreting s-inconsistent items, when participants come up with a situation model that renders 'seeing' or 'being aware of' an object compatible with its being behind the agent (who may look into a mirror, turn the head around, or – for 'aware' – hear or feel the object behind, or just know all along where it is). Where sentences are deemed very implausible (1), participants arguably fail to come up with such a situation model (58% of the time for s-inconsistent 'see'-sentences with visual objects, 18% for their 'aware'-counterparts). Where participants give higher ratings we take them to assess how probable the situation they envisage (e.g., Chuck sees a spot on the wall behind him, in a mirror) is to actually arise, under the circumstances inferable from the sentence context.

For s-inconsistent 'see' sentences with epistemic objects (like 'Kelly sees the possibilities from which she has turned away'), viable interpretation also requires suppression of spatial conclusions, as it involves purely epistemic interpretation of 'see' and purely temporal interpretation of the post-object sequel. On such a purely metaphorical interpretation (*Kelly knows what possibilities she did not take up*), s-inconsistent items are deemed distinctly plausible and *more* plausible than s-consistent counterparts (Appendix A). Even so, s-inconsistent 'see'-sentences with epistemic objects ('Kelly…') were rated neutral, and *less* plausible than their s-consistent counterparts. By the reasoning that motivated our plausibility predictions (Sec. 3.1), we infer that in the epistemic conditions, spatial conclusions from 'see' were not completely suppressed from the situation model, as the strongly associated vision schema was comprehensively retained for interpretation, and prevented a purely metaphorical reading.

By contrast, the fact that context does not affect the plausibility of 'aware'-sentences with epistemic objects suggests that participants managed to obtain a purely epistemic interpretation of such 'aware'-items. I.e.: If 'aware' initially activates the vision schema, participants completely

suppress its irrelevant (non-epistemic) component features. This seems plausible, given 'aware' is less strongly associated with that schema than 'see'. In summary, plausibility ratings for sentences with epistemic objects suggest that contextually irrelevant spatial components of the vision schema get completely suppressed in judging 'aware'-, but not 'see'-sentences.

Let's finally consider a potential confound. Some observed responses might be due not to inappropriate spatial inferences but to temporally indexed existential inferences, which are perfectly appropriate: S-inconsistent items like 'Jack sees the problems *he left behind*' use a post-verbal cancellation phrase that implies the patient (*problem*) no longer exists, for the agent (*Jack*). But also in its epistemic sense, 'see' implies the simultaneous existence of patient and agent (you can only 'see my point', as and when I have one). Hence, the objection goes, the clash of an *appropriate* inference from the verb with an implication of the cancellation phrase may have driven the lower plausibility of s-inconsistent items with epistemic objects.

Half of these items employed a post-verbal context implying non-existence ('Jack…'), and half the contexts implied existence (e.g., 'Kelly sees the possibilities *from which she has turned away*' suggests the possibilities are still there, though Kelly no longer takes them into account). This lets us examine the possible interaction between different stereotypical implications from the verb and the cancellation phrase. The objection implies lower plausibility ratings for the nonexistence-implying 'see'-items ('Jack…') than the existence-implying ones ('Kelly…'). We analysed responses for items with existence- and nonexistence-implying cancellation phrases separately (Figure 4; for details see Appendix D).

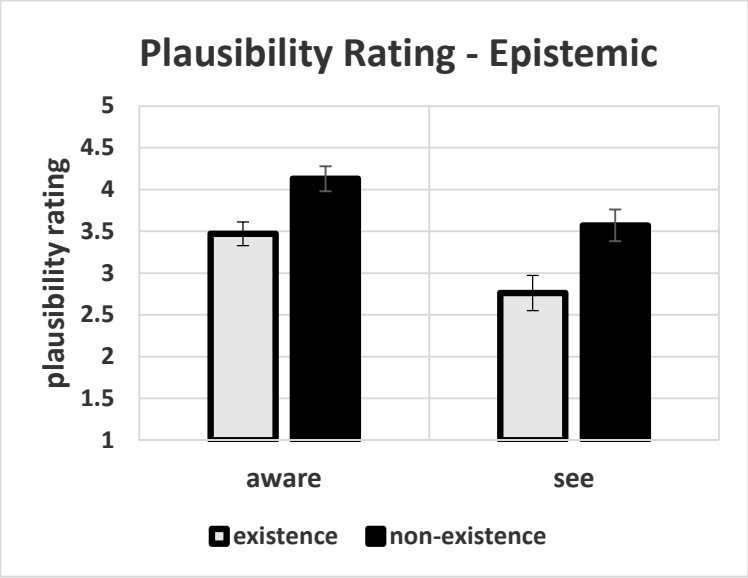

Figure 4. Plausibility ratings for s-inconsistent items with epistemic objects where cancellation phrases imply existence vs. non-existence. Error bars show the standard error of the mean.

Existence implications had the opposite effect than that predicted by the objection: Items with nonexistence-implying cancellation phrases were deemed *more* plausible than items with existence-implying cancellations; the former were deemed plausible, the latter not.

We interpret these plausibility ratings in line with our Salience Bias Hypothesis (SBH), and as illustrating the relevance of condition (iii) of the SBH: Inappropriate inferences go through where contextually irrelevant schema components get only partially suppressed, and suppression remains partial only where initial activation of all components of the dominant situation schema

is complemented – as per condition (iii) – by lateral cross-activation from frequently co-occurring schema components which are contextually relevant. In the vision schema, the directional component *X is in front of S* regularly co-occurs with – indeed, is dependent upon – the component *X exists at the same time as S*. In interpreting epistemic uses, the former can therefore be completely suppressed where the latter is ruled out as contextually irrelevant, but not where the existential component is confirmed as contextually relevant. This component is ruled out as contextually irrelevant by our nonexistence-implying cancellation phrases. Hence both regular companions can be completely suppressed, and no inappropriate spatial inference goes through to affect plausibility judgments. By contrast, the contextual relevance of the existential component is confirmed by our existence-implying cancellation phrases. Hence its regular spatial companion cannot be completely suppressed, and an inappropriate inference goes through and leads to lower plausibility ratings.

This explanation needs to be tested against our pupil data. It implies great suppression effort (with merely partial success) for 'see'-items with existence-implying cancellations and less effort (with more success) for 'see'-items with nonexistence-implying cancellations. This predicts significant pupil dilations at least for items with existence-implying cancellations, and larger pupil responses for them than for nonexistence-implying items. We therefore analysed pupil responses for items with existence- and nonexistence-implying cancellations separately (Figure 5; see Appendix D).

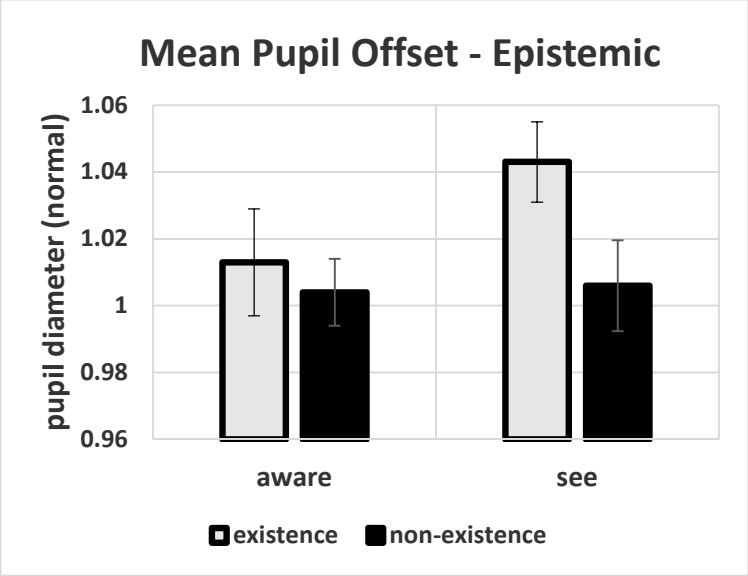

Figure 5. Baseline adjusted pupil diameter in 1000ms time window after sentence offset, for s-inconsistent items where cancellation phrases imply existence vs. non-existence. Error bars show the standard error of the mean.

The findings are consistent with our predictions and support the proposed explanation.

The comparison of responses to 'see'- and 'aware'-items, finally, illustrates the relevance of pronounced salience imbalances – as per condition (i) of SBH – for the thus aptly named 'salience bias': Similar patterns in pupil results (and in eye-tracking results; Fischer and Engelhardt, 2019) for 'see' and 'aware' suggest that the vision-schema is retained for interpreting epistemic uses also of 'aware' – satisfying condition (ii) of SBH. Since 'aware' is associated less strongly with the vision schema, contextually inappropriate schema components are easier to suppress completely. Accordingly, complete suppression of irrelevant components is possible even when – as per (iii) – these receive lateral co-activation from regular companions that are contextually relevant: Our

participants regarded 'aware'-items even with existence-implying cancellation phrases still as distinctly plausible. This suggests that, even in unhelpfully phrased contexts, the use of the Retention/Suppression Strategy to interpret less salient uses of polysemous words only leads to inappropriate inferences where these words display pronounced salience imbalances and have a dominant sense far more salient than the others.

4. Conclusion: Main Findings and Philosophical Relevance

4.1 Main Findings

By combining pupillometry with plausibility ratings, our study provided evidence of inappropriate spatial inferences from epistemic uses of the verb 'see', which influence further cognition. Building on previous studies that used only output measures to study stereotypical inferences from perception- and appearance verbs (Fischer and Engelhardt 2016; 2017a; Fischer et al. 2019), it provides more rigorous support for the proposed Salience Bias Hypothesis (SBH): Case descriptions prompt inappropriate stereotypical inferences that influence further cognition when (i) they employ familiar polysemes (e.g., 'see') which have a clearly dominant sense, in a less salient sense, to describe cases which deviate from the dominant (e.g., *seeing*) stereotype. Inappropriate inferences licensed only by the dominant sense then occur when (ii) these descriptions are processed by retaining and partially suppressing the dominant stereotype (e.g., the *vision* schema), rather than by activating a distinct stereotype (e.g., a *knowledge* schema). I.e.: Inappropriate inferences occur where case-descriptions are processed as descriptions of unusual cases of one thing (*seeing*), rather than cases of something else. Inappropriate inferences go through when (iii) the cases described diverge from some, but not all, the central features of the stereotype.

Our experiments (pre- and main study) show such inferences are made even from less salient uses which are perfectly familiar (like the epistemic use of 'see' illustrated by 'I see your point'), and influence further thought even when competent speakers know the inferences at issue are inappropriate, namely, typically lead to conclusions that fail to be true (see Sec. 2.3). Future research could fruitfully examine whether this salience bias results from a system design that strikes the best balance between processing effort and accuracy of information inferred and retained, given real-world task demands (*cf.* Ferreira and Lowder 2016).

Future research should also examine to what extent the findings of our experiments with undergraduate students carry over to expert philosophers (*cf.* Nado 2014; Machery 2017, ch.5). Our findings suggest that neither familiarity with the relevant less salient senses nor high levels of insight into the inappropriateness of the inferences prevents the inappropriate inferences predicted by the Salience Bias Hypothesis. To render philosophers less prone to the inferences of interest, other factors would need to set them apart from our participants.[10]

[10] We regard reflectiveness as the most relevant of the factors discussed in the metaphilosophical literature. Livengood et al. (2010) found that philosophers are especially reflective, as assessed by the Cognitive Reflection Test (Frederick 2005); that is, they are more likely to reflectively check intuitive judgments generated by automatic inferences. This tendency may render thinkers less vulnerable to endorsing explicit conclusions of inferences which they know to be inappropriate. Even then, however, it is unlikely to mitigate the covert influence of implicit conclusions that are tacitly presupposed in further judgment or reasoning (see Sec. 4.2). Studies that operationalised reflectiveness in a variety of ways found that higher reflectiveness does not mitigate susceptibility of intuitive case judgments to clearly irrelevant factors which affect judgment beyond participants' awareness (Colaço et al. ms; Weinberg et al. 2012).

4.2 Philosophical Relevance

Assuming that stereotypical comprehension inferences are ubiquitous, they will be occurring continuously when thinkers read or hear verbal case descriptions in philosophical arguments and thought experiments. Since the activation processes in semantic memory that support them occur in both language comprehension and production (Pickering and Garrod 2013, Stephens et al. 2010), these inferences also occur in the sub-vocalised soliloquy characteristic of much philosophical thought (*cf.* Carruthers 2002).

The salience bias is then liable to assert itself where philosophers give special, less salient uses to words which already have dominant uses in ordinary discourse, and the associated dominant schemas are functional for interpreting less salient uses. While only psycholinguistic research will reveal whether dominant schemas are thus functional, we can readily observe that philosophers often give rare uses to familiar words to meet specific needs of philosophical reflection or to talk about unusual cases. We presently consider two examples to clarify how contextually inappropriate stereotypical inferences from such rare uses may vitiate philosophical arguments and thought experiments, respectively.

To date, the 'Sources Project' in experimental philosophy (see Fn.1) focussed on how automatic cognitive processes directly generate intuitive judgments that are explicitly endorsed, and sought to assess the evidentiary value of these intuitions by considering the reliability of the underlying processes. But stereotypical inferences can influence judgments about verbally described cases also indirectly, by leading to implicit conclusions that enter into the comprehender's tacit situation model on which subsequent judgment and reasoning about the case are based (Author et al. 2019; Saint-Germier ms.). The present study showed that contextually inappropriate stereotypical inferences can contribute to such situation models. Such inferences lead to contextually defeated conclusions and create inconsistencies in situation models. Their exposure facilitates further epistemological assessment of case verdicts based on those models, which may complement assessments of process reliability.

We now discuss two examples where, we submit, stereotypical inferences from rarefied uses of familiar words led to contextually defeated implicit conclusions that entered situation models and were assumed by case verdicts. Our first example is the 'argument from hallucination' against naïve realism about perception. In its historically most influential version, it argues for the existence of mind-dependent objects of perception ('sense-data'), which separate subjects from any physical objects in their environment (Ayer 1956/1990, 90, Fish 2010, 12-15, Macpherson 2013, 12-13, Smith 2002, 194-197). Philosophers of perception often use familiar appearance- and perception-verbs in a rare 'phenomenal' sense, which lacks existential, factive and spatial implications (e.g., Ayer 1956/1990, 90, Fish 2010, 6, Jackson 1977, 33-49, Macpherson 2013, 5; *cf.* Chisholm 1957, 44-48); they do so when they wish to merely describe perceivers' subjective experience and to talk about genuinely unusual cases, like hallucinations. The following classic statement of the argument from hallucination 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 1956/1990, 90).

The second half of the argument then generalises from this special case (hallucination) to all cases of visual perception.

We propose a fresh diagnostic analysis of the argument which identifies a 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' = S has an experience like that /as of seeing an F), he therefore *can* be said to 'see a physical dagger' – but 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.

Deploying the Salience Bias Hypothesis, we suggest the argument is driven by stereotypical inferences from the special phenomenal use of 'see' in (1), to implicit conclusions (like 2 below) that are defeated by contextual information (like 3 below) – and are tacitly assumed, even so, in further reasoning (here: from explicit premises (1) and (3) to conclusion (4)):

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

On some interpretations of (1), Macbeth hallucinates a purely imaginary dagger; on others, he hallucinates the actual murder weapon he will use (which just isn't around at this point), and (2) and (3) quantify over locally restricted domains, with the second 'there' understood as 'before Macbeth'. In the former case, we submit, the argument involves parallel existential and spatial inferences from (1); in the latter case, the argument relies exclusively on the kind of spatial inferences our experiment examined, now from the least salient use of 'see'.[11] Either way, the inference from (1) to (2) would be licensed by the dominant visual sense of 'see', but is not licensed by the special phenomenal sense, which lacks factive implications about the subject's environment and creates an intensional context not admitting of quantification (Forbes 2013); either way, the inference leads to a conclusion that is inconsistent with contextual information (3, true by assumption). 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'.

We submit the Salience Bias Hypothesis can explain why competent thinkers make this bad inference and assume its contextually defeated conclusion in further reasoning: (i) 'see' is used here in a special phenomenal sense significantly less salient even than the epistemic sense employed in our experiment (see Table 1). We suggest (ii) the dominant word schema is retained and deployed for its interpretation (*cf.* Giora 2003, Giora et al. 2014): A situation-model that instantiates this schema with specific patient-role fillers (e.g., *dagger*) 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 (Ortony 1993, Searle 1993). However, (iii) what it is like to see something is strongly associated with most other

[11] Recall (from Sec. 2.3) that our findings about spatial inferences from epistemic uses of 'see' are set to carry over to uses that are even less salient and inferences that are less transparently inappropriate.

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, Wilson 2002). Accordingly, it is hard to retain only the phenomenal component and suppress the schema's other central components (indeed, this should be even harder than to retain only the less fully integrated epistemic component). If the conditions under which stereotypical enrichment is vulnerable to salience bias – (i)-(iii) in the Salience Bias Hypothesis – thus obtain, our account can explain why stereotypical inferences licensed by the dominant visual sense are made also inappropriately, and despite explicit marking (Sec. 2.2), from phenomenal uses to contextually defeated conclusions that are assumed in further reasoning.

Turning to directions for future work, we finally consider a prominent thought experiment which uses verbal case descriptions as intuition pumps. David Chalmers (1996, 94-96) famously uses esoteric 'zombie cases' to generate modal insights: He applies the word 'zombie' (in the new technical sense of 'phenomenal zombie') to imaginary beings which deviate from the zombie stereotype – lifeless face, rigid stare, mechanical movements, etc. – by physically being and behaving just like us. The Salience Bias Hypothesis suggests that if the stereotype associated with the arguably dominant 'Hollywood-sense' of 'zombie' is deployed to interpret Chalmers' utterances, judgments about the imagined physical duplicates of ours will implicitly assume they have those stereotypical features – even though the stereotype-deviation is made explicit. Those features suggest the duplicates sporting them express no feelings and manifest no sensations. Unlike the explicit assumption that the duplicate creatures are and behave just like us, the implicit attribution of those features would therefore seem compatible with the judgment that the creatures lack conscious experience. When prompted by Chalmers' case descriptions, the modal intuition that zombies which are and behave just like us but lack conscious experience are 'obviously' possible is arguably based on those contextually defeated implicit attributions. A follow-up study will therefore consider the hypotheses that Chalmers' zombie case descriptions trigger contextually inappropriate inferences to attributions of stereotypical zombie-features and that these attributions are instrumental for generating the modal intuition.

4.3 Metaphilosophical Conclusion and Methodological Consequences

At the metaphilosophical level, the Salience Bias Hypothesis provides us with an empirically supported and more specific version of the prominently discussed *'esotericity thesis'*. This previously speculative thesis states that intuitive judgment about 'esoteric' cases is less reliable than about 'normal' cases (Weinberg 2007; 2015; Williamson 2007). The Salience Bias Hypothesis identifies specific conditions – (i)-(iii) above – under which contextually inappropriate automatic inferences are made from verbal case descriptions and affect further judgment and reasoning. While the relevant differences between 'esoteric' and 'normal' cases have previously remained elusive (Cappelen 2012, Weinberg 2015), the present hypothesis suggests a key difference lies in divergence from vs. conformity with stereotypes associated *with the words employed to describe* the cases: The Salience Bias Hypothesis suggests that judgments and inferences arising, more specifically, from the ubiquitous process of stereotypical enrichment are, respectively, less likely to be true or less reliable when they are about cases that are 'unusual' in deviating in some central respects from the 'dominant stereotypes' associated with those words' dominant senses.[12]

[12] Empirical work on other cognitive processes may identify lapses in reliability for further kinds of 'unusual' cases.

The range of cases that are 'unusual' in this sense extends far beyond the cases commonly thought of as 'esoteric' (perfect hallucinations, zombies, fake barns, etc.). In particular, they include normal instantiations of another category (e.g., unusual cases of 'seeing' may be standard cases of *knowing*). If so, error can easily be avoided by paraphrasing the case description. However, this simple remedy is not so readily available for cases that are 'genuinely unusual' in that they do not fit any of our familiar lexicalised categories or stereotypes; in such cases, we typically resort to extending the use of familiar words by applying them to these cases – even though the cases diverge from the associated stereotypes in some central respects. Furthermore, thinkers tend to know little about such genuinely unusual cases (like hallucinations or physical duplicates); therefore, conclusions of the inappropriate stereotypical inferences then predicted by the Salience Bias Hypothesis are yet less likely to be suppressed through integration with background knowledge (*cf.* Metusalem et al. 2012, Fischer and Engelhardt 2017b), and yet more likely to go through. While verbal descriptions even of perfectly ordinary cases are liable to prompt contextually inappropriate stereotypical inferences when involving less salient uses of familiar words, inappropriate inferences about genuinely unusual cases are therefore particularly hard to avoid and particularly liable to influence further judgments and reasoning.

Whereas the initial esotericity thesis basically suggested that philosophers should refrain from treating intuitions about unusual cases as evidence, the Salience Bias Hypothesis has more specific and productive methodological consequences that leave some scope for philosophical use of even genuinely unusual cases. The Salience Bias Hypothesis suggests that the problem ultimately arises not from the unusual nature of cases *per se* (which 'merely' exacerbates the problem), but from extending the use of words with clearly dominant senses, namely, in ways that lead to the dominant schemas being deployed in interpreting descriptions of cases that deviate from some central schema features. As long as we manage to describe an unusual case without stretching the use of familiar words in this way, the risk of inappropriate stereotypical inferences influencing philosophical verdicts about the case is low. Psycholinguistic norming work (below) can provide us with the necessary insights into salience structures and processing strategies. The Salience Bias Hypothesis thus motivates the following methodological suggestions.

First, explicit marking of less salient word-uses (Givoni et al. 2013) and of deviations from dominant stereotypes (Garrett and Harnish 2007) need not prevent inappropriate stereotypical inferences from occurring and influencing judgment and reasoning (Sec. 2.2). Wherever possible, we should therefore replace problematic words ('see', 'zombies') by expressions without contextually inappropriate stereotypical implications ('hallucinate', 'physical duplicates'). Often, simple listing and sentence-completion tasks suffice to elicit stereotypical implications (McRae et al. 1997).

Second, where we have to use a word that has inappropriate stereotypical associations in one of its senses, we should ensure that this is not a dominant sense clearly more salient than all others. We can determine salience differences between different senses, e.g. of 'know' (Hansen et al., under review), by collecting frequency information (e.g. from *WordNet* or fresh corpus studies) and using sentence-completion tasks to determine prototypicality (Chang 1986).

Third, where we cannot avoid using a word which has a dominant sense, with contextually inappropriate stereotypical implications carried by that sense, it is still an open question whether the relevant less salient use will be interpreted by deploying the stereotype associated with that dominant sense – or will activate a distinct schema (*cf.* Sec. 2.2). We can then use priming (Giora

and Fein 1999) as well as reading-time measurements with self-paced windows (Giora et al. 2007a) or eye-tracking (Fischer and Engelhardt, 2019), to determine whether the dominant stereotype is retained and deployed for interpretation of the relevant less salient use. Only where verbal ingenuity and psycholinguistic norming fail to yield suitable case descriptions do philosophers need to do without verdicts about the unusual case described. We suspect, however, that properly developed vignettes devoid of problematic expressions (which speak, e.g., of 'duplicates', not 'zombies') will often prompt different verdicts than currently popular case descriptions.

4.4 Conclusion

The identification of a salience bias which asserts itself under philosophically relevant conditions lets us identify past mistakes and avoid future error. To be epistemologically sound, any philosophical practice that employs verbally described cases in thought experiments and arguments, and accords some default evidentiary status to spontaneous judgments or conclusions about these cases, will need to take precautions, including psycholinguistic norming of case-descriptions, wherever descriptions apply familiar words to unusual cases.[13]

References

Alexander, J., Gonnerman, C., & Waterman, J. (2015). Salience, and epistemic egocentrism. In J. Beebe (ed.), *Advances in Experimental Epistemology* (pp. 97-118). London: Bloomsbury.

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

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.

Bordalo, P., Coffman, K., Gennaioli, N., Shleifer, A. (2016). Stereotypes. *Quarterly Journal of Economics*, 131, 1753-1794.

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

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.

Brown, P.M. & Dell, G.S. (1987). Adapting production to comprehension: The explicit mention of instruments. *Cognitive Psychology 19,* 441–472.

[13] Both authors contributed equally to material development and interpretation of results. Paul Engelhardt conducted data gathering and analysis. Eugen Fischer undertook the remaining research. For helpful comments on earlier drafts and closely related material the authors are indebted to the Editor in Charge and two anonymous reviewers for this journal, to Kathryn Francis, Max de Gaynesford, Rachel Giora, James Hampton, Nat Hansen, Aurélie Herbelot, and Joachim Horvath, and to audiences at the 8th International Conference of the Experimental Philosophy Group UK (Norwich, July 2017), the 2nd Conference of the Experimental Philosophy Group Germany (Osnabrück, Nov. 2017), the 2018 meeting of the European Epistemology Network (Amsterdam, June 2018), and at guest lectures at the Universities of Jena (Dec. 2016) and Reading (Nov.2017).

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

Carruthers, P. (2002). The cognitive functions of language. *Behavioral and Brain Sciences*, 25, 657–74

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

Chalmers, D. (1996). *The Conscious Mind*. Oxford: OUP

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

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

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

Colaço, D., Kneer, M., Alexander, J., & Machery, E. (ms). *On second thought: An assessment of the reflection defense*.

Elman J.L. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognition*, 33, 547–582.

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.

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

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

Fein, O., Yeari, M., & Giora, R. (2015). On the priority of salience-based interpretations: the case of sarcastic irony. *Intercultural Pragmatics,* 12, 1-32.

Ferreira, F., & Lowder, M.W. (2016). Prediction, information structure, and good-enough language processing. *Psychology of Learning and Motivation*, 65, 217-247

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. (2014). Philosophical Intuitions, Heuristics, and Metaphors. *Synthese*, 191, 569-606

Fischer, E., & Engelhardt, P.E. (2016). Intuitions' linguistic sources: Stereotypes, intuitions, and illusions. *Mind and Language*, *31*, 67–103.

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

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

Fischer, E., & Engelhardt, P.E. (2019). Eyes as windows to minds: Psycholinguistics for experimental philosophy. In E. Fischer & M. Curtis (eds.), *Methodological Advances in Experimental Philosophy* (pp. 43–100). Bloomsbury.

Fischer, E., Engelhardt, P.E., Horvath, J., & Ohtani, H. (2019). Experimental Ordinary Language Philosophy: A Cross-Linguistic Study of Defeasible Default Inferences. *Synthese*. DOI 10.1007/s11229-019-02081-4

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

Forbes, G. (2013). Intensional transitive verbs. In E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. Fall 2013. http://plato.stanford.edu/archives/fall2013/entries/intensional-trans-verbs/

Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*, *19*, 25–42.

Garrett, M., & Harnish, R.M. (2007). Experimental pragmatics: testing for implicatures. *Pragmatics and Cognition*, 17, 245-262.

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. *Noûs*, 50, 133-164.

Gibbs, R.W. (1994). *The Poetics of Mind. Figurative Thought, Language, and Understanding*. Cambridge: CUP

Gibbs, R.W. (2002). A new look at literal meaning in understanding what is said and implicated. *Journal of Pragmatics*, 34, 457–486.

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

Givoni, S., Giora, R., & Bergerbest, D. 2013: How speakers alert addressees to multiple meanings, *Journal of Pragmatics*, 48, 29-40.

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

Hansen, N., Porter, J.D., & Francis, K. (under review). A corpus study of "know": On the verification of philosophers' frequency claims about language.

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

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.

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

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

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

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

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

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

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. & Wilson, A. (2001). *Word Frequencies in Written and Spoken English*. London: Longman.

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

Livengood, J., Sytsma, J., Feltz, A., Scheines, R., & Machery, E. (2010). Philosophical Temperament. *Philosophical Psychology*, *23*, 313–330.

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

Machery, E. (2017). *Philosophy within its proper bounds*. Oxford: Oxford University Press.

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.

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.

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

O'Brien, E.J., & Albrecht, J.E. (1992). Comprehension strategies in the development of a mental model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18, 777-784.

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.

Pollock, J. (1986). *Contemporary Theories of Knowledge*. Savage, MD: Rowman and Littlefield.

Pust, J. (2017). Intuition. In E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, http://plato.stanford.edu/ archives/sum2017/entries/intuition/

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

Rosch, E. (1975). Cognitive reference points. *Cognitive Psychology,* 7, 532-547.

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.

Saint-Germier, P. (ms). *Getting Gettier straight: thought experiments, deviant realization, and pragmatic enrichment.*

Searle, J. (1993). Metaphor. In A. Ortony (ed.), *Metaphor and Thought*, 2nd ed. (pp. 83-111). Cambridge: CUP.

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.

Sloman, S.A., Love, B.C. & Ahn, W.K. (1998). Feature centrality and conceptual coherence. *Cognitive Science*, 22, 189–228

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

Starmans C., & Friedman O. (2012). The folk conception of knowledge. *Cognition*, 124, 272-83.

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.

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.

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.

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.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science*, 185, 1124-1131

Vu, H., Kellas, G., Metcalf, K., & Herman, R. (2000). The influence of global discourse on lexical ambiguity resolution. *Memory and Cognition*, 28, 236–252.

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

Weinberg. J. (2015). Humans as instruments, or 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, and J. Hawthorne (eds.), *Oxford Handbook of Philosophical Methodology* (pp. 287-308). Oxford: OUP.

Weinberg, J. M., Alexander, J., Gonnerman, C., & Reuter, S. (2012). Restrictionism and reflection: Challenge deflected, or simply redirected? *The Monist*, *95*, 200–222.

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.

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

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review*, 9, 625-636.

Appendix

Section A – Item development: plausibility of metaphorical interpretations

26 undergraduate psychology students from the University of East Anglia, all native speakers of English, participated for participant pool credits. They read paraphrases capturing purely metaphorical interpretations of the 24 critical items with epistemic objects used in the main study. Paraphrases replaced 'see' by 'know' and resolved spatial time metaphors in sequels. Paraphrase of s-consistent ('that lie ahead') and s-inconsistent ('that lie behind') critical items yielded future- and past-directed items in equal number, e.g.:

1. Joe knows what problems he will have in the future. (future-directed)
2. Jack knows what problems he had in the past. (past-directed)

24 fillers were either future- or past-directed; half of each kind were designed to be plausible, half implausible. Participants rated each sentence on a 1-to-5 plausibility scale. Paraphrases of s-inconsistent critical items (like 2) were rated significantly higher than paraphrases of s-consistent critical items (like 1) $t1(25)=-13.07$, $p<.001$, $\eta^2=.87$; $t2(11)=-2.79$, $p<.05$, $\eta^2=.41$. The means were 4.03 (SD=.37) and 2.85 (SD=.43), respectively. One sample *t*-tests with test-value 3 revealed the former paraphrases (like 2) were deemed distinctly plausible (significantly >3; $t(25)=14.21$, $p<.001$), the latter (like 1) neutral (not significantly different from 3; $t(25)=-1.75$, $p>.10$).

These findings suggests that, in the main study, spatial inferences from 'see' prevented purely metaphorical interpretation, in *all* epistemic items: past-directed paraphrases were deemed plausible, while s-inconsistent originals were not, *and* s-consistent originals were deemed plausible, while paraphrases directed at the hard-to-predict future were not.

Section B – Main study: materials

Our main study employed 48 critical sentences as well as 48 fillers and two practice items. A female native speaker of British English recorded all items. Items employed the same epistemic patient nouns as in the pre-study (Section 2.3). The visual and epistemic objects had similar mean frequencies (M=86, SD=83 and M=79, SD=79, respectively, based on Leech et al. 2001).

Epistemic

1. Joe is aware of/sees the problems that lie ahead.
2. Jack is aware of/sees the problems he left behind.
3. Nelly is aware of/sees the hardship that lies ahead.
4. Claire is aware of/sees the hardship she left behind.
5. John is aware of/sees the opportunities that lie ahead.
6. Jim is aware of/sees the opportunities he left behind.
7. Bob is aware of/sees the commitments that lie ahead.
8. Bill is aware of/sees the commitments he left behind.
9. The high-flier is aware of/sees the factors that stand in his way.
10. The high-flier is aware of/sees the factors that push him onwards.
11. Emma is aware of/sees the challenges facing her.
12. Ellie is aware of/sees the challenges she has overcome.
13. Josh is aware of/sees the issues facing him.
14. Jeb is aware of/sees the issues he has overcome.
15. The colonel is aware of/sees the trouble brewing at his front.
16. The general is aware of/sees the trouble brewing in his rear.
17. Mary is aware of/sees the possibilities that have popped up before her.
18. Kelly is aware of/sees the possibilities from which she has turned away.
19. Carl is aware of/sees the risks that have popped up before him.
20. Kurt is aware of/sees the risks from which he has turned away.

21. Clara is aware of/sees the solutions that have popped up before her.
22. Chloe is aware of/sees the solutions from which she has turned away.
23. Jeff is aware of/sees the options that have popped up before him.
24. Jeff is aware of/sees the options from which he has turned away.

Visual

25. Matt is aware of/sees the spot on the wall facing him.
26. Chuck is aware of/sees the spot on the wall behind him.
27. Carol is aware of/sees the picture on the wall facing her.
28. Sheryl is aware of/sees the picture on the wall behind her.
29. Max is aware of/sees the old knocker on the door facing him.
30. Tim is aware of/sees the old knocker on the door behind him.
31. Mona is aware of/sees the car on the road facing her.
32. Laura is aware of/sees the car on the road behind her.
33. Alan is aware of/sees the shoppers waiting in front of him in the queue.
34. Alex is aware of/sees the shoppers waiting behind him in the queue.
35. The hiker is aware of/sees the friend walking in front of him.
36. The rambler is aware of/sees the friend walking behind him.
37. Nora is aware of/sees the colleague seated in front of her.
38. Ellie is aware of/sees the colleague seated behind her.
39. Larry is aware of/sees the toddler playing in front of him.
40. Jerry is aware of/sees the toddler playing behind him.
41. Harry is aware of/sees the scene unfolding before his eyes.
42. Sam is aware of/sees the scene unfolding at his back.
43. Fay is aware of/sees the accident happening before her eyes.
44. Dawn is aware of/sees the accident happening at her back.
45. Fred is aware of/sees the musicians performing before her eyes.
46. Mark is aware of/sees the musicians performing at her back.
47. Amy is aware of/sees the quarrel unfolding before her eyes.
48. Ann is aware of/sees the quarrel unfolding at her back.

In hindsight, we noticed two epistemic items (viz. 15-16) had slipped in whose spatial sequel does not admit of metaphorical (non-spatial) interpretation. Examination of results from these two potentially problematic items showed that neither was a statistically significant outlier, and running the analyses with these items excluded from all analyses (plausibility and pupillometry) revealed no change to any of the reported patterns or significance in the data. We therefore retained the items in the analysis reported in the main text.

Section C – Main study: procedure and analysis

Participants were seated at the eye tracker (SR Research EyeLink 1000) and instructed verbally. They placed their chins on a chinrest. After a 9-point calibration procedure, participants completed two practice trials and 96 experimental trials. Each participant saw 'see'- and 'aware'-versions of critical items in equal number, in each condition, as verbs and context-phrases were rotated across lists in a Latin-Square design. On each trial, a fixation cross appeared for 1500ms prior to sentence onset. The sentence was played out on computer speakers, and after sentence offset the fixation cross remained on the screen for 1000ms. 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. Endpoints were explained as 'very implausible' (1) and 'very plausible' (5), and the midpoint (3) as 'neither plausible nor implausible; the decision feels arbitrary'. Mean pupil diameter was measured during the second half of the sentence and during the offset time period. 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 increasing or decreasing compared with the immediately preceding time window. For plausibility and pupil diameter, we defined outliers as means +/-3 SDs from the mean. Outliers were replaced with the mean of that condition, which avoids listwise deletion and subsequent loss of power. There was one outlier in the plausibility results and three outliers in pupil results (<1% of the data). Analyses were conducted with subjects (*F*1) and items (*F*2) as random effects.

Section D – Main study: analysis of existence vs. non-existence-implying cancellations

Two annotators independently classified the cancellation phrases used in epistemic items and judged that 's/he left behind' and 's/he has overcome' (used in six of twelve relevant items) imply non-existence and that the remaining phrases imply existence.

Plausibility ratings: A 2×2 (verb × cancellation) ANOVA revealed significant main effects of verb $F(1,36)=13.13$, $p<.01$, $\eta^2=.15$ and cancellation $F(1,36)=17.85$, $p<.001$, $\eta^2=.33$. (Since this analysis was based on a post-hoc classification of items, not all participants received equal numbers of items in each of the four present conditions as items were rotated across lists. Therefore, this and the following ANOVA treated the cancellation variable as between-subjects – the more conservative approach.) Consistent with previous findings, 'aware'-items were deemed more plausible than 'see'-items. Crucially, items with nonexistence-implying cancellation phrases were deemed more plausible than items with existence-implying cancellations. 'See'-items with nonexistence-implying cancellations (mean 3.57) were judged more plausible than existence-implying counterparts (2.76) (confirmed by an exploratory paired-samples t-test $t(34)=-2.85$, $p<.01$, $\eta^2=.18$). Indeed, 'see'-items with nonexistence-implying cancellations were deemed distinctly plausible (*sic*) (mean >3, $t(18)=3.00$, $p<.01$), whereas 'see'-items with existence-implying cancellations fell into the lower plausibility category (neutral, not significantly different from 3, $t(18)=-1.13$, $p=.27$). While 'aware' items with nonexistence-implying cancellations were deemed more plausible than existence-implying counterparts (Figure 4), both were deemed distinctly plausible (mean >3) $t(18)=7.42$, $p<.001$ and $t(18)=3.34$, $p<.01$, respectively.

Pupil results: A 2×2 (verb × cancellation) ANOVA showed a trend of cancellation $F(1,36)=2.85$, $p=.09$, $\eta^2=.04$ in the predicted direction. This trend can be traced to a marginally larger pupil response for 'see'-items with existence- than with nonexistence-implying cancellations $t(34)=2.00$, $p=.053$, $\eta^2=.10$, as assessed with a paired-samples t-test. Moreover, 'see'-items with existence-implying cancellations prompted significant pupil dilations $t(18)=3.59$, $p<.01$ as assessed with a one-sample t-test, while none of the other conditions did (p's >.43).