Final draft version (July 2, 2022) of chapter in D. Bordonaba (ed.), *Experimental Philosophy of Language*. New York: Springer

How understanding shapes reasoning: Experimental argument analysis with methods from psycholinguistics and computational linguistics

Eugen Fischer and Aurélie Herbelot

Abstract Empirical insights into language processing have a philosophical relevance that extends well beyond philosophical questions about language. This chapter will discuss this wider relevance: We will consider how experimental philosophers can examine language processing in order to address questions in several different areas of philosophy. To do so, we will present the emerging research program of experimental argument analysis (EAA) that examines how automatic language processing shapes verbal reasoning – including philosophical arguments. The evidential strand of experimental philosophy uses mainly questionnaire-based methods to assess the evidentiary value of intuitive judgments that are adduced as evidence for philosophical theories and as premises for philosophical arguments. Extending this prominent strand of experimental philosophy, EAA underpins such assessments, extends the scope of the assessments, and expands the range of the empirical methods employed: EAA examines how automatic inferences that are continually made in language comprehension and production shape verbal reasoning, and draws on findings about comprehension biases that affect the contextualisation of such default inferences, in order to explain and expose fallacies. It deploys findings to assess premises *and* inferences from premises to conclusions, in philosophical arguments. To do so, it adapts methods from psycholinguistics and recruits methods from computational linguistics.

1. Experimental argument analysis: motivation and key ideas

Our chapter will present the emerging research program of experimental argument analysis.[1] The present section will introduce the motivation and guiding ideas of EAA by outlining how the research program emerged from empirical engagement with the critical strand of Oxford ordinary language philosophy. Sections 2-3 will present a relevant example: a body of research that (i) documents a previously unrecognised comprehension bias that affects the processing of polysemous words (i.e., words with several distinct but related senses) and (ii) deploys findings to explain and expose a fallacy of equivocation in a key argument from the philosophy of perception (viz., the argument 'from hallucination'). Sections 4-5 will explain the empirical methods employed. In this way, the chapter will illustrate how experimental philosophers can study language processing to address issues from different areas of philosophy.

As practiced in the mid-20th century, ordinary language philosophy (OLP) was analytic philosophy's first attempt to overcome limitations of armchair reflection through the use of (informal) experiments (Hansen and Chemla 2015), (peer-based) focus groups (Urmson 1969), and empirical surveys (Murphy 2014). This makes OLP an important historical precursor of current experimental philosophy.[2] OLP's 'critical' strand popularized the idea that many

[1] The program's proponents – who include the present authors – have cheekily appropriated a simple but broad label for a quite specific research program that focuses on how automatic default inferences from words influence verbal reasoning. Important work that does not share this specific focus but could well also be described as 'experimental argument analysis' includes work on fallacies in reasoning with conditionals (Pfeifer 2012; Pfeifer and Tulkki 2017, *cf.* Skovgaard-Olsen et al. 2019) and with metaphors (e.g., Ervas et al. 2015; 2018).

[2] For critical discussion of this claim, see Longworth (2018). For helpful discussion of how corpus methods can be used to empirically implement OLP, see Sytsma et al. (this volume).

characteristically philosophical problems arise from conceptual confusions or verbal fallacies and can be 'dissolved' by exposing these confusions or fallacies in the underlying arguments (e.g., Austin 1962; Ryle 1949, 1954; Waismann 1968; for a review, see Schroeder 2006). *Prima facie*, this idea makes most sense where philosophical problems are generally regarded as arising from antinomies, like sceptical problems and the problems of free will, mental causation, and perception. As typically conceived, these problems arise from persuasive arguments that lead to conclusions that appear to rule out familiar facts, as recognised by common sense (*cf.* Fischer 2011; Papineau 2009). The resulting problems are often articulated by questions that ask how familiar facts are as much as possible. They occasion the kind of wonder that Plato notoriously regarded as the starting-point of all philosophizing (*Theaetetus* 155b-d). Let's call them 'Platonic puzzles'.

A case in point is the 'problem of perception' (Smith 2002). As typically conceived (for a review, see Crane and French 2021), it arises from an antinomy developed by arguments 'from illusion' and 'from hallucination'. These arguments proceed from the uncontroversial assumptions that illusions and hallucinations occur (or, in more cautious versions, that these phenomena are at least possible). In a first step, the arguments conclude that in the cases considered – i.e., illusions or hallucinations – viewers are aware, or directly aware, of subjective and immaterial objects (perceptions or sense-data) in their minds, rather than physical objects in their environment (see below, Sect. 2). In a second step, the arguments then generalise to all cases of visual perception. These arguments challenge what philosophers regard as the common-sense view of vision, which grants viewers direct access to physical objects, without detour via any immaterial objects of sight. The arguments raise the problem 'that if illusions and hallucinations are possible, then perception, as we ordinarily understand it, is impossible' (Crane and French 2021, §2; *cf.* Hume 1777) and motivate the question at the centre of many debates about the nature of perception: How is perception, as we ordinarily understand it, even possible? (Robinson 1994; Smith 2002)

If conceptual confusions or verbal fallacies prevent the underlying arguments from getting off the ground, this question is ill-motivated and needs to be rejected rather than answered. It is hence plausible to try to 'dissolve' problems of this kind by exposing such fallacies in the underlying arguments. J.L. Austin sought to 'dissolve' the problem of perception by exposing 'seductive (mainly verbal) fallacies' that act as 'concealed motives' for formulating the problem (Austin 1962, p.5). That the fallacious inferences are 'concealed' means, on a charitable interpretation, that thinkers are not conscious of making the inferences and presupposing their conclusions, in the relevant arguments (Fischer 2014). Those inferences are *automatic*, i.e., they require no attention, are unconscious, and insensitive to the thinkers' goals (Bargh et al. 2012; Evans and Stanovich 2013). Austin sought to clarify 'the root ideas behind the uses' of key words that are employed in the targeted arguments (Austin 1962, p.37). He went on to discuss inferences which are supported by those 'root ideas' but have subtle contextual defeaters (Austin 1962, pp. 37-43). He seems to suggest that the targeted arguments involve automatic inferences which go through even where they are defeated by context.

While multiple further applications offer themselves, efforts to develop EAA are animated by an interest in 'dissolving' philosophical problems like the problem of perception and draw on psycholinguistic findings to develop Austin's suggestions (Fischer and Engelhardt 2016, 2017a, 2017b, 2020; Fischer, Engelhardt and Sytsma 2021; Fischer et al. 2021).

EAA identifies what Austin called 'root ideas behind the uses of words' as stereotypes associated with words. As standardly conceived, *stereotypes* are implicit knowledge structures

in semantic memory that encode information about statistical regularities observed in the physical or discourse environment (e.g., tomatoes are typically red and juicy) (McRae and Jones 2013). They thus capture what psychologists call 'world knowledge' and philosophers regard as empirical knowledge. They encode statistical information about typical and diagnostic properties of category members, which may be objects, people, or events (Hampton 2006). Complex stereotypes (known as 'situation schemas') encode information about typical features of events or actions, agents, 'patients' acted on, and typical relations between them (Ferretti et al. 2001; Hare et al. 2009; McRae et al. 1997). This knowledge about the world plays a key role in language processing (Elman 2009): Stereotypes can be associated with individual nouns and verbs. These words (like 'tomato') activate stereotypical information rapidly (within 250ms) (for a review, see Engelhardt and Ferreira 2016), automatically, and largely irrespective of context, i.e., *'by default'* (Machery 2015).[3] Activated stereotypes support *stereotypical inferences* to attributions of stereotypical features (the tomato talked about will be red) (Levinson 2000). These automatic inferences are unavoidable, get things right more often than not, but are defeasible ('the tomato was still green'). The crucial Austinian suggestion then becomes that the targeted philosophical arguments involve defeasible stereotypical inferences that are contextually defeated – but whose conclusions are presupposed in further reasoning, anyway.

Since verbal stimuli trigger these inferences by default, these inferences provide the first materials from which language users construct the situation model, i.e., the representation of the situation talked about that provides the basis for further reasoning about that situation (Zwaan 2016). Default inferences are therefore bound to shape verbal reasoning profoundly. Indeed, Levinson (2000, p.28) suggested these inferences are instrumental in facilitating effective communication in the face of the 'articulation bottleneck': Pre-articulation processes in speech production are 3-4 times faster than normal speech (Wheeldon and Levelt 1995), as are parsing processes and comprehension inferences in speech comprehension (Mehler et al. 1993). Default inferences that deploy our statistical knowledge about the world allow hearers to rapidly fill in detail. Anticipating such inferences allows speakers to skip mention of typical features and use fewer words. Default inferences thus facilitate effective communication.

The first key idea EAA derives from Austin is to examine how default inferences shape verbal reasoning – for a start, in philosophical arguments and with a view to resolving Platonic puzzles. Considerable care is required, however, to develop Austin's more specific suggestion that contextually inappropriate default inferences might be an important source of reasoning fallacies. Psycholinguistic work on sentence comprehension suggests that language users are good at contextualising default information. For a start, nouns and verbs together ('The mechanic checked…') can swiftly activate complex stereotypes that encode information about recurrent situations (car inspections) and are not activated by individual words on their own (Bicknell et al. 2010; Matsuki et al. 2011). Activation for the less specific stereotypes initially activated by individual words then decays where they lack contextual support (Oden and Spira 1983) and the contextually more appropriate schema enters into the situation model.

In a neo-Gricean framework, stereotypical inferences have accordingly been conceptualised as governed by a heuristic, Levinson's (2000) I-heuristic, that tells hearers that,

[3] That information is 'activated' means that it is made more readily available for use in further cognitive processes. Information activated by a verbal stimulus thereby becomes more readily available for processes ranging from word recognition (e.g., recognising the next word) to sentence parsing (e.g., assigning thematic roles like agent and patient) and verbal reasoning.

in the absence of explicit indications to the contrary, they should assume that the situation talked about conforms to the relevant stereotypes, and should treat *the most specific stereotypes* activated (say, about car inspections) as the most relevant. Moreover, stereotypical inferences that clash with contextual information or background knowledge can be suppressed within one second (Fischer and Engelhardt 2017b; *cf.* Faust and Gernsbacher 1996).

Diagnosing fallacies in philosophical arguments requires caution at the best of times. The interpretation of philosophical texts is governed by widely accepted principles of charity. These principles tell us to credit authors with linguistic competence and rationality. This requirement creates a tension with the attribution of fallacies to authors (Adler 1994; Lewinski 2012). Medium-strength principles of charity resolve the tension by allowing interpreters to attribute fallacies to authors only if the attribution is backed up by an empirically supported explanation that explains when and why even competent thinkers commit fallacies of the relevant kind (Thagard and Nisbett 1983). Given that competent language users are generally good at contextualising default information, the Austinian suggestion that influential philosophical arguments rely on contextually inappropriate stereotypical inferences is in particularly acute need of such an explanation.

Philosophical argument analysis is often regarded as the epitome of an armchair activity. However, the Austinian suggestion that fallacious automatic inferences drive philosophical arguments requires empirical support. The need for empirical support arises from the facts that the posited inferences are fallacious and automatic. An a priori reconstruction of fallacious verbal reasoning can only specify inference chains that *could* have led thinkers from a premise to a conclusion it does not entail. Thinkers have no privileged access to automatic inferences. Their self-reports or acceptance of a proposed reconstruction therefore cannot provide a justified answer to the question of which inference chain – of many potentially relevant chains – *actually* led them from premise to conclusion. To support the hypothesis that a particular automatic inference drives an argument, we need to document the posited inference experimentally. Moreover, the attribution of fallacious inferences to competent thinkers like philosophers is constrained by principles of charity that ask interpreters to support such attributions with empirical error theories that explain when and why such fallacies occur. Hence, we need experimental evidence not only of the specific inferences posited, but also for accounts that explain them.

Inspired by these sources, EAA examines how default inferences that go on continually in language comprehension and production drive verbal reasoning. It focuses on how stereotypical inferences shape philosophical arguments and seeks to expose contextually inappropriate stereotypical inferences in such arguments. This requires developing psycholinguistic explanations of these fallacies and conducting experiments (i) to examine these explanations and (ii) to document the specific fallacies posited in arguments.

EAA seeks to explain why inappropriate stereotypical inferences influence further reasoning by reference to comprehension biases. We now present the approach through a case study on the argument from hallucination: We present a reconstruction of the argument that takes it to rely on contextually cancelled stereotypical inferences from polysemous perception verbs (Sect. 2). We then outline a psycholinguistic explanation of when and why such inferences are made from polysemous words (Sect. 3). We finally explain how hypotheses about inappropriate stereotypical inferences have been examined with methods from psycholinguistics (Sect. 4) and computational linguistics (Sect. 5).

2. Example: a philosophical argument

Consider a classic statement of the argument from hallucination, by the influential British mid-20th century philosopher A.J. Ayer. This statement carefully distinguishes between a perceptual sense of the verb 'to see' and a phenomenal sense that serves purely to describe the viewer's subjective experience and thus lacks all factive, spatial, etc., implications:

> '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, p.90, bold added).

The second half of the argument then generalises from this special case to all cases of visual perception (Macpherson 2013; Smith 2002). The argument is commonly intended as a deductive argument. The following reconstruction remains as close to the text as possible and builds a deductive argument from the bits highlighted in bold above (explicit assumptions and conclusions numbered in round brackets, implicit assumptions in square brackets):

(1) 'There was no [real] dagger there.'
(2) 'Macbeth did <u>see</u> a dagger.'

To deductively infer that Macbeth did not see a real dagger ('But still not a real dagger'), we need an implicit assumption:

[3] If Macbeth **saw** a real dagger, there was a real dagger there. By (1) & [3] with modus tollens:
(4) 'Macbeth did not **see** a real dagger.'
[5] Macbeth did not **see** any other physical object, either. By (4) & [5]:
(6) 'Macbeth did not **see** a physical object.' Hence:
(7) 'If Macbeth **saw** any object, he saw a non-physical object, i.e., "sense-datum."'[4] By (7) & (2):
(8) 'Macbeth saw a sense-datum.'

This reconstruction posits a previously little noted fallacy of equivocation: The implicit assumption [3] uses 'see' in the perceptual sense that has factive implications – if S sees an F (say, a dagger), then an F is there. Hence the conclusions derived from it, directly or indirectly, need to use the verb in the same perceptual sense (highlighted in bold). This includes (7). But Ayer then derives the crucial conclusion (8) from (7) and (2) – even though (2) explicitly uses the verb in the phenomenal sense (underlined) that lacks factive implications. *Pace* (3), that Macbeth 'saw' a real dagger in this sense does not imply there was a real dagger. While this criticism applies regardless of the specific explication of the phenomenal sense used, the following illustration may help to bring out the fallacy. On one interpretation of Ayer's explanation of the phenomenal sense (Fischer and Engelhardt 2020), 'S $\text{sees}_{\text{PHEN}}$ an F' means 'S has an experience like that of seeing an F'. Macbeth is meant to have an experience just like

[4] Arguably, this step assumes a dichotomous distinction between 'physical objects' and 'sense-data', whereby any non-physical object of vision is a private sense-datum.

that of seeing a physical dagger. In the phenomenal sense, he can therefore be said to 'see a physical dagger', because that is exactly what his experience is like. In this phenomenal sense, he can*not* be said, e.g., to see a translucent non-physical dagger (his experience is not like that). In Ayer's text, the move from 'Macbeth saw a dagger' (in the phenomenal sense) to 'but still not a real dagger' is hence fallacious.

This reconstruction faces the challenge from the principle of charity: Ayer explains the two senses of perception verbs, before setting out the argument, and flags their uses, in the argument. Our reconstruction suggests he made an inference from the phenomenal use that is licensed only by the perceptual sense. This violates Ayer's own explanation of the phenomenal sense, i.e., a self-imposed semantic rule. Analytic philosophers are competent speakers. Our reconstruction thus implies that a competent speaker violated a semantic rule he explained himself a few lines up, in an inference from a premise where the special use of the word was explicitly marked. The principle of charity hence requires us to explain why such a competent thinker would commit the relevant fallacy under the circumstances. We explain the fallacy of equivocation by reference to a comprehension bias that occurs in polysemy processing. This bias asserts itself under conditions that frequently arise in philosophical reflection.

3. Example: a comprehension bias

Polysemes activate a unitary representation of semantic information that is deployed to interpret utterances which use the word in different senses (Macgregor, Bouwsema and Klepousniotou 2015; Pylkkänen, Llinás and Murphy 2006). The findings we reviewed above (Sect. 1) about how words cue world knowledge for rapid deployment in utterance interpretation suggest a unitary representation is typically built around stereotypes associated with the word. Different senses can sometimes be generated by rules (as in metonymy) and sometimes not (as in metaphor). In the latter case, of 'irregular polysemy', the unitary representation consists in overlapping clusters of features (Brocher, Foraker and Koenig 2016; Klepousniotou et al. 2012), and may include overlapping stereotypes.

Different components of these unitary representations get activated in different strength by the verbal stimulus. The stimulus activates the features shared by related senses quickly and strongly, regardless of context. By contrast, the activation of unshared features is a function of their relative exposure frequency (Brocher et al. 2018): The more often the language user encounters the word in one sense, rather than another, the more strongly the (unshared) features associated with (only) that sense are activated, when the user encounters the word. This is consistent with a sensible predictive strategy: The use frequencies observed to date provide the baseline probability that the word is being used in this sense, on this occasion. This baseline activation may be boosted by context (op. cit.). Another factor influencing strength of activation is prototypicality: Features deemed to make for particularly good examples of the relevant category are activated more rapidly and strongly (Hampton 2006). Strength of activation thus depends on *linguistic 'salience'* (Giora 2003). Unlike the contextual salience involved in familiar salience biases (see Taylor and Fiske 1978 for a review), this is not a contextual magnitude, but a function of relative exposure frequency over time modulated by prototypicality.

Interpreting any particular use of a polyseme then requires activating all contextually relevant, but unshared features, and suppressing all contextually irrelevant, but activated features. Consider a simple case where the features relevant for interpreting a subordinate use are a subset of the features that make up the stereotype associated with the dominant sense: the

verb ‘to see’ is associated with a situation schema (the ‘*seeing*-schema’) that includes the typical agent features *S has eyes*, *S looks at X*, *S knows X is there*, and *S knows what X is*; patients typically are medium-sized dry goods; and typical relations between patients and agents include *X is in front of S* and *X is near S*. To interpret the purely epistemic use illustrated by ‘Jack saw Jane’s point’, precisely the last two agent features are relevant: Jack knows there is a point of Jane’s and he knows what it is. These need to be retained, while the other features need to be suppressed, applying the ‘*Retention/Suppression strategy’* (Giora 2003).

Two circumstances may prevent complete suppression of contextually irrelevant features: First, suppose features irrelevant for the subordinate sense (as, e.g., *X is in front of S* is irrelevant for the epistemic sense of ‘see’) are associated with a clearly dominant sense (e.g., the visual sense of ‘see’) that is far more frequent than all other senses. Then these irrelevant features will receive very strong initial activation (Brocher et al. 2018). Second, frequently co-instantiated component features of a stereotype exchange lateral co-activation (Hare et al. 2009; McRae et. al. 2005). Where only some, but not all of the components associated with the dominant stereotype are relevant for interpreting a subordinate use, the contextually relevant features will continue to pass on activation to the contextually irrelevant features. Where these two factors come together, strong initial activation of contextually irrelevant features is followed by their continued cross-activation. This makes complete suppression impossible. When merely partially suppressed, irrelevant features continue to support stereotypical inferences.

This creates a *linguistic salience bias* (Fischer and Engelhardt 2019, 2020; Fischer and Sytsma 2021): When

i. one sense of an irregular polyseme is much more salient than all others,
ii. interpretation of utterances with a subordinate sense requires suppression of features associated with that dominant sense, and
iii. some, but not all, of the features strongly associated with the dominant sense are contextually relevant

then

1. contextually inappropriate stereotypical inferences supported by the dominant sense will be triggered by the subordinate use as well, and
2. these automatic inferences will influence further judgment and reasoning.

I.e.: When an irregular polyseme is seriously unbalanced and the Retention/Suppression strategy is used to interpret subordinate uses, even competent thinkers cannot help being influenced by automatic inferences that are cancelled by contextual information. Thinkers are then swept along by defeasible inferences, even when these are defeated by the context.

The relevant conditions are often met in philosophy: Philosophers often give special but related uses to familiar words that have clearly dominant senses from ordinary discourse. Arguably, the use of such polysemes is an important source of fallacies in philosophical reasoning. Studies to date provided evidence that linguistic salience bias affects inferences from subordinate uses of perception verbs (Fischer and Engelhardt 2017a, 2017b, 2019, 2020; Fischer, Engelhardt and Herbelot 2022), from phenomenal uses of appearance verbs that are involved in arguments from illusion (Fischer and Engelhardt 2016; Fischer, Engelhardt and Sytsma 2021; Fischer et al. 2021), from philosophical uses of ‘zombie’ (Fischer and Sytsma 2021), and from purely descriptive uses of the verb ‘to cause’ in morally valenced cases (Livengood, Sytsma and Rose 2017; Livengood and Sytsma 2020). Crucially, a study with academic philosophers revealed that they are no less susceptible to the bias than laypeople

(psychology undergraduates) (Fischer, Engelhardt and Herbelot 2022). This finding allows to invoke linguistic salience bias to explain fallacies of equivocation in philosophical arguments.

4. Methods from psycholinguistics

Most of these studies have adapted the *cancellation paradigm* that psycholinguists developed to study automatic comprehension inferences. In this paradigm, participants read or hear sentences where the expression of interest is followed by text that defeats or 'cancels' the inference that is by hypothesis triggered by that expression. To examine, for example, whether participants make automatic inferences from 'S sees X' to *X is in front of S* we can ask them to read sentences like:

Sheryl sees the picture on the wall behind her.

If the inference is made, the resulting clash of the conclusion with the sequel causes comprehension difficulties which require cognitive effort to overcome. When we expend cognitive effort, our pupils dilate (Kahneman 1973; Laeng et al. 2012). When we struggle to integrate new information with information inferred from previous text, we need longer to read the cancellation phrase (e.g., 'behind her') and make more backwards eye movements from that phrase (Patson and Warren 2010). Finally, perceived conflicts prompt signature electrophysiological responses ('N400s') (Kutas and Federmeier 2011). These 'online' measures (which tap into cognitive processes as they unfold) can be used to examine whether specific automatic inferences are triggered by words, as people read or hear them.

As noted above, however, initially activated stereotypical information may simply decay in the absence of contextual support (Oden and Spira 1983), and stereotypical inferences that clash with contextual information or background knowledge can be suppressed within one second (Fischer and Engelhardt 2017b; *cf.* Faust and Gernsbacher 1996). Either way, initially triggered automatic inferences fail to influence further judgment and reasoning. To study whether automatic inferences influence further cognition, we therefore complement online measures with subsequent plausibility ratings: Where inferences are not suppressed, perceived clashes with sequels will persist and lead to lower ratings.

This paradigm is illustrated by three studies on spatial inferences from subordinate uses of perception verbs 'see' and 'aware of' (Fischer and Engelhardt 2017b, 2019, 2020). Prior corpus analyses revealed that the purely epistemic sense ('I see your point') is the most salient of the subordinate senses of 'see'. In one paradigmatic study, occurrence frequencies in a random 1000-sentence sample from the *British National Corpus* (BNC) served as proxy measure for exposure frequency, and frequencies from a sentence completion task measured prototypicality (see Table 1). A pre-study revealed that members of our participant pool reject spatial inferences from purely epistemic uses yet more strongly than spatial inferences from other subordinate uses (like the phenomenal use) (Fischer and Engelhardt 2020). Our studies therefore considered spatial inferences from purely epistemic uses of 'see'.

Table 1. Occurrence and completion frequencies for 'see' (from Fischer and Engelhardt 2020)

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

We now consider in some detail the fixation-times study (Fischer and Engelhardt 2019) that demonstrates the most subtle methodology that allows us to examine both automatic inferences and the mechanism of polysemy processing. In reading, the eye moves in stops and starts. Readers tend to fixate most, but not all words, as their eyes skip the contextually most predictable words *and* move backwards at points of difficulty, so that many words get fixated repeatedly. Difficulties at different stages of text comprehension then manifest themselves in different eye-tracking measures (Clifton et al. 2016; Rayner et al. 2004). Difficulties in word recognition depend on the word's frequency, length, and predictability in local context. These jointly determine the first-pass fixation time for the word. By contrast, difficulties in integrating local interpretations of a few adjacent words into comprehensive interpretations of an entire sentence show in late measures: They lead to longer second-pass or total reading times, and to more regressions to earlier text. In the cancellation paradigm, inferences from the verb phrase lead to integration difficulties and longer total reading times for conflict regions or source regions from which problematic inferences originate.

We used sentences with distinct verb, object, and context regions (see Table 2), and a 2×2×2 within-subject design: Verb regions employed either 'see' or the contrast verb 'is aware of', which is less frequently used with concrete objects which we become aware of in virtue of seeing them, and is therefore less strongly associated with the *seeing*-schema than 'see'. Object regions employed either concrete objects (like 'the picture') or abstract objects (like 'the problems') which invite a purely epistemic interpretation of the previous verb. Context regions were either consistent or inconsistent with the *seeing*-schema, that is, '*s-consistent'* or '*s-inconsistent*', as they placed objects either before or behind the agent – literally, for concrete objects, and metaphorically, for abstract objects. For epistemic items, the purely epistemic sense of 'see' ('know/understand something') and familiar spatial time metaphors (whereby ahead = in the future; behind = in the past) facilitate purely metaphorical interpretations (e.g., *Joe knows what problems he had in the past*). S-inconsistent epistemic items used three different cancellation phrases: 'lie behind', 'has overcome', and 'has turned from'.

Table 2. Example stimuli and regions of interest for eye movement analysis (from Fischer and Engelhardt 2019)

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

Linguistic salience bias asserts itself – only – where pronounced salience imbalances go with use of the Retention/Suppression strategy (Sect. 3). When an ambiguous word with a dominant meaning is used in a less salient sense and disambiguated by immediate post-verbal

context (as in our items), this increases first-pass reading times on the disambiguating region (Sereno et al. 2006). If the Retention/Suppression strategy is used, however, also late (second-pass, and total) reading times for the object region will be longer for 'see'-sentences with epistemic than with visual objects, across all 'see'-items. To determine whether purely epistemic uses of 'see' are interpreted by retaining and selectively suppressing features from the *seeing*-schema, we measured the late (second-pass and total), reading times for the object region – and, as predicted, found them longer for epistemic than for visual objects.

This allowed us to put the linguistic bias hypothesis directly to the test. To document spatial inferences from 'see', we measured total reading times for s-consistent and s-inconsistent context regions. Unsurprisingly, in sentences with visual objects, inviting a literal interpretation of 'see', reading times were higher for s-inconsistent context regions. As predicted, however, the same held true for sentences where abstract objects invited epistemic interpretation of the verb. This suggests that also epistemic uses of 'see' prompted spatial inferences from 'S sees X' to *X is before S*.

But would these inferences influence further cognition – or be suppressed right away? To address this crucial question, we considered plausibility ratings. If spatial inferences are completely suppressed, participants win through to a purely metaphorical interpretation. On this interpretation, items with 'see' and 'aware of' mean exactly the same, namely, e.g., that *Joe knows what problems he had in the past* or *what problems he will have in the future*. Moreover, in a norming study, participants rated the plausibility of such paraphrases that made purely metaphorical interpretations explicit – and deemed past-facing items more plausible than future-facing items (Fischer and Engelhardt 2020). Complete suppression therefore predicts that s-inconsistent epistemic items with 'see' and 'aware of' will be deemed equally plausible, and more plausible than s-consistent items ('that lie ahead of him', which are future-facing on a metaphorical interpretation).

By contrast, incomplete suppression will yield the opposite pattern: The space–time metaphors in our items 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, Imai and Boroditsky 2002). Persistent spatial inferences from 'see' will prevent purely metaphorical interpretation also of the space-time metaphors, engage spatial reasoning, and create the impression of a conflict, specifically in s-*in*consistent 'see'-items. Prevention of purely metaphorical interpretation can result in persistent 'visual' interpretation that identifies, e.g., the problems seen with visible objects (*Mountaineer Joe sees the difficult-to-cross crevice that lies behind him*). Even where the object (say, problem) is not identified as visual, the impression of a conflict between spatial implications from 'see' and the sequel will make s-inconsistent 'see'-sentences feel 'weird' and *less* plausible than s-consistent 'see'-sentences. Linguistic salience bias predicts suppression difficulties for 'see', but not the contrast verb 'aware of'. The predicted difficulties will hence lead participants to find s-inconsistent epistemic items with 'see' less plausible than counterparts with 'aware'.

Findings clearly decided in favor of the linguistic salience bias hypothesis (see Figure 1). All predicted plausibility differences were significant (p's$\leq$.001) and translated into categorical differences: Whereas 'aware'-sentences with epistemic objects were deemed distinctly plausible (mean ratings significantly above mid-point '3' of our scale), regardless of whether they had a s-consistent or s-inconsistent sequel, 'see' sentences with epistemic objects were judged distinctly plausible with s-consistent sequels but neutral (means no different from '3') with s-inconsistent contexts.

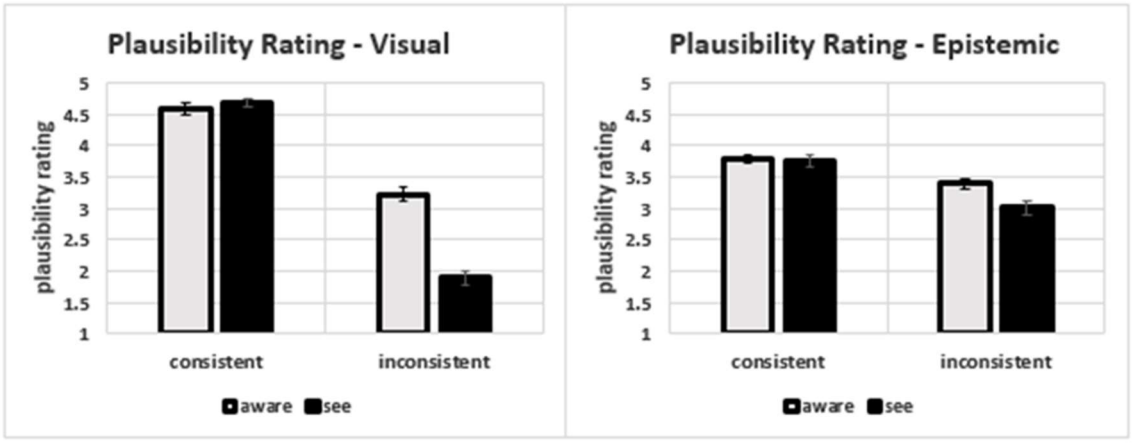

Figure 1. Mean plausibility ratings for each of the eight conditions in the eye tracking study. Error bars show the standard error of the mean. (From Fischer and Engelhardt 2019)

Our initial reconstruction of the argument from hallucination posited at its root contextually cancelled factive inferences from phenomenal uses of perception verbs (Sect. 2). Linguistic salience bias could explain such inferences (Sect. 3). Inferences of interest can be documented by implementing the psycholinguistic cancellation paradigm with a combination of online measures (such as fixation times) and plausibility ratings (Sect. 4). Extant studies provide evidence of the bias by documenting, among other things, spatial inferences from epistemic uses of 'see'. To complete its treatment of the target argument, EAA needs to further document the specific inferences posited as its source, namely, factive inferences from *phenomenal* uses of 'see' and 'be aware of'. These two lexical items are most frequently used to state the argument. Both have clearly dominant uses – visual and epistemic, respectively – with factive implications that are cancelled by standard explanations of the phenomenal sense and by contextual information in the argument from hallucination. Fixation-time studies to document these inappropriate inferences are ongoing.

5. Methods from computational linguistics

To study default interpretations and inferences, we can complement behavioural experiments with computational models which allow us, for example, to test whether the patterns present in a specific type of linguistic data could in principle account for the decisions made by actual human participants when presented with a given task. Modern natural language processing (NLP) models are based on machine learning techniques: given some input, a particular task to perform, and some light supervision (i.e., feedback on correctness or accuracy of performance), these models will teach themselves to recognize patterns of interest in the data, and use such patterns to perform the task with the highest possible accuracy. This pipeline is ideal to simulate any behaviour that is learned from exposure to a particular type of experience – such as the encounter of words in particular kinds of contexts, over time.

One particularly useful paradigm for the study of linguistic salience and default interpretations is known as 'Distributional Semantics' (DS) (Erk 2012; Lenci 2018).[5] DS is a computational technique akin to methods in corpus linguistics, but suited to the analysis of big data. It is inspired by Wittgenstein's (1953) suggestion that questions of meaning should be

[5] For helpful discussion of DS in this volume, see the chapter by Grindrod.

rendered tractable as questions about the use of words and proceeds from an analysis of how frequently words are used together: In line with the maxim that 'you shall know a word by the company it keeps' (Firth 1957, p.11), DS is used to characterise lexical items in terms of their distributional similarity in a corpus (i.e., in terms of the extent to which they co-occur in the corpus with the same other words, in the same proportions). Distributional similarity is then taken as a measure of semantic similarity. As input, a DS model receives a large corpus in a language of interest. As output, it provides semantic representations of lexical items in the form of points, or 'vectors', in a multidimensional space. The creation of semantic vectors is understood as defining the meaning of lexical items with respect to a finite number of properties, which correspond to the dimensions of the space. Whenever we apply further mathematical operations to the original vectors, we 'move' them through space and thereby emphasize or reduce the importance of a property *with respect to a particular question.*

To generate its output, a traditional statistical model computes the number of co-occurrences of a target word with other words or constituents in its context of use, and applies a number of operations to the resulting counts, with the aim of bringing the space to the best possible geometric configuration. What counts as an adequate configuration depends on the task at hand. It is, however, widely accepted that, at the very least, the space should provide good agreement with human judgements of semantic similarity. Such judgments are usually elicited through behavioural tasks where participants rate the comparative relatedness of concepts ('is a cat more similar to an elephant or to a motorbike?'). The system encodes its space in a way that reflects human intuitions (the 'cat' vector is closer to 'elephant' than to 'motorbike'). The resulting vector space is then amenable to further processing using the tools of linear algebra.

In recent years, the original DS paradigm has evolved from a method based on corpus statistics to a set of neural network techniques designed to simulate word prediction in the context of an utterance (Mikolov et al. 2013). This facilitates the use of DS to study the linguistic salience of word senses: Some of the new 'predictive' systems provide designs that standardly output contextualised representations of lexical items (Devlin et al. 2019; Peters et al. 2017). That is, given a sentence, they will return a point for each word in the sentence, reflecting the sense modulations that the word may have undergone in that particular context. In this revised paradigm, a lexical item is not associated with a single point representing its type but rather with one or more clusters of vectors encoding the *instances* of the type. By analysing the set of vectors formed by a word type across a large corpus, we find the dominant sense of that type, encoded in the largest clusters, and have an opportunity to compute its linguistic salience.

We can also apply further machine learning techniques to the clusters, in order to simulate human performance on more specific linguistic tasks. Crucially, this allows us to study the salience of uses that are not defined in terms of the original clusters. For example, Fischer, Engelhardt and Herbelot (2022) were interested in the relative frequency of perceptual and non-perceptual uses of 'see' and 'be aware of', to gauge how often the need for suppression of spatial inferences from these verbs is likely to arise. Clusters of vectors corresponding to instances of the two verbs were used to supervise a system to classify instances as perceptual or non-perceptual uses. The classifier generates judgements in line with human subjects. This allows for deployment over hundreds of thousands of new instances from different corpora, to assess the respective occurrence frequencies of the two usages, in the different kinds of discourse reflected in the different corpora, without human intervention beyond the annotation

of initial training samples.

DS models also allow us to address natural follow-up questions. Fischer, Engelhardt and Herbelot (2022) used them to address two questions about the relevance of expertise: (1) Classifiers are validated by assessing their verdicts against human annotations and showing that they perform better than a simple chance heuristic (which classifies all occurrences of a word as instances of its dominant use in the corpus). Successful validation, with major improvements on this baseline, indicates that context words (without even syntactic parsing) provide enough information to identify the different uses of interest (e.g., non-perceptual uses of 'see' and 'aware of'). No expert knowledge seems required. (2) Cross-domain classification lets us determine whether differences in linguistic diet might make it difficult for people to identify uses of interest in unfamiliar discourse settings. Fischer and colleagues trained their classifier on a sample from one corpus (e.g., the BNC) and tested its accuracy on an annotated sample from another corpus (e.g., a specialist philosophy of perception corpus). Major improvements over baseline suggested that no specialist knowledge is required to identify uses of interest (e.g., non-perceptual vs perceptual uses of 'see' and 'aware of').

We now turn from linguistic salience to default inferences. Most NLP systems are predictive at some level, and thus perform default inferences with respect to the particular linguistic phenomenon they have been trained for. A traditional probabilistic architecture performs prediction by assigning a probability distribution to a range of alternatives. For instance, given an instance of a word type in the context of a specific utterance, a Word Sense Disambiguation model assigns different probabilities to the different senses of that word. One of those senses can usually be regarded as its default interpretation, in that its probability tends to be the highest across usages in different utterances. This default interpretation would be the one activated when the word is presented outside of context. But even in context, a system may incorrectly assign a default sense to a particular usage if the surrounding utterance is not strong enough to defeat the bias learned by the model. Similarly, a so-called 'language model', i.e., a system that is simply trained to predict the next word in a sentence, has a probability distribution over all words in its vocabulary, and should learn that the most likely continuation of the verb 'thank' at the beginning of a sentence is the pronoun 'you'. Neural networks are not probabilistic *per se*, but they implement functions that simulate probability distributions and are often interpreted as such. So most systems encode some notion of 'preference' or 'default' over a set of given alternatives for a particular phenomenon of interest. It follows that if a computational model encodes defaults as (quasi-) probability distributions, and if it has a modular internal structure mirroring specific theoretical choices (as it usually does), then it should be possible to implement modules that test different aspects of default inference that are not directly 'visible' in psycholinguistic work.

In particular, computational models allow us to study complex internal interactions which involve both world knowledge and linguistic knowledge. Consider the inferences supported by event knowledge, in the form of learned 'schemas' or 'scripts'. The rapid activation of such world knowledge by combinations of verbs and nouns facilitates quite specific inferences (Bicknell et al. 2010; Matsuki et al. 2011). For example, the verb 'cut' evokes different instruments depending on its context of use: 'cut the cake' implies a kitchen knife while 'cut the grass' is associated with lawnmowers. Such implicit inferences are made transparent in probabilistic systems involving Frame Semantic Parsing (Das et al. 2014). In such systems, conceptual objects such as categories and event schemas are represented in a lexicon of 'frames' (Fillmore 1982), which encode some default semantic structure for the

concept: for instance, the concept CUT might encode that some underspecified instrument is necessary for the agent to complete the action denoted by the lexical item. A frame semantic parser is a type of system which learns to interpret sentences as the composition of frames, using knowledge of the most likely conceptual associations given an utterance context. The world knowledge activated by the sentence is thus made explicit, and the generic process by which defaults are invoked and/or cancelled is formalised as probabilistic reasoning.

Computational models can also be used to study the interaction of defaults across levels of linguistic knowledge, spanning the morphological, syntactic, semantic and pragmatic properties of a sentence. Erk and Herbelot (2021), for instance, formalise sentence interpretation as the process of giving probabilities to semantic structures made of situations, concepts and semantic roles. The model makes it possible to explicitly describe the conflicting constraints between global and local utterance context. For example, in a sentence such as *Alice lost her wallet at the bank while fishing*, the (global) fishing scenario cancels the (local) lexical interpretation of 'bank'. It also accounts for cases where senses genuinely oscillate between different interpretations, as in classic puns (e.g., *The astronomer married the star*, where 'star' conserves some ambiguity).

In summary, humans are good at contextualising default information (Sect. 1). Even so, contextualisation sometimes fails. This raises questions about the conditions for such failures. To address such questions, computational models can complement psycholinguistic experiments. In particular, such models make it possible to ask which constraints from one linguistic level may undo the defaults at a different level, and to test the hypothesis, given exposure to a certain type of data.

6. Conclusion

Language processing is a key topic in the philosophy of language. Experimental philosophers can, however, study it empirically to address philosophical questions and problems that extend well beyond the philosophy of language. We considered an emerging research program that does so: Extending the evidential strand of experimental philosophy (Machery 2017; Mallon 2016), experimental argument analysis (EAA) seeks to explain and expose fallacies in verbal reasoning, starting with philosophical arguments that generate characteristically philosophical problems ('Platonic puzzles') – in several areas of philosophy. To do so, EAA examines how default inferences that automatically go on in language comprehension and production shape verbal reasoning. Extant research in the program focuses on contextually cancelled stereotypical inferences and explains such bad inferences as resulting from comprehension biases like the linguistic salience bias that arises in polysemy processing. We presented first findings as well as methods from psycholinguistics and computational linguistics that can be employed for the purpose. These include a novel implementation of the psycholinguistic cancellation paradigm as well as distributional semantic models.

References

Adler, Jonathan. 1994. Fallacies and alternative interpretations. *Australasian Journal of Philosophy* 72: 271-282. https://doi.org/10.1080/00048409412346091.

Austin, John Langshaw 1962. *Sense and Sensibilia*. Oxford: Oxford University Press.

Ayer, Alfred Jules 1990 [1956]. *The Problem of Knowledge*. London: Penguin.

Bargh, J.A., K.L. Schwader, S.E. Hailey, R.L. Dryer, and E.J. Boothby. 2012. Automaticity in social-cognitive processes. *Trends in Cognitive Sciences* 16: 593-605. https://doi.org/10.1016/j.tics.2012.10.002.

Bicknell, K., J.L. Elman, M. Hare, K. McRae, and M. Kutas. 2010. Effects of event knowledge in processing verbal arguments. *Journal of Memory and Language* 63: 489–505. https://doi.org/10.1016/j.jml.2010.08.004.

Boroditsky, L., and M. Ramscar. 2002. The roles of body and mind in abstract thought. *Psychological Science* 13: 185-188. https://doi.org/10.1111%2F1467-9280.00434.

Bottini, R., D. Crepaldi, D. Casasanto, V. Crollen, and O. Collington. 2015. Space and time in the sighted and blind. *Cognition* 141: 67–72. https://doi.org/10.1016/j.cognition.2015.04.004.

Brocher, A., S. Foraker, and J.-P. Koenig. 2016. Processing of irregular polysemes in sentence reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 42(11): 1798–1813. https://psycnet.apa.org/doi/10.1037/xlm0000271.

Brocher, A., J.-P. Koenig, G. Mauner, and S. Foraker. 2018. About sharing and commitment: the retrieval of biased and balanced irregular polysemes. *Language, Cognition and Neuroscience* 33(4): 443-466. https://doi.org/10.1080/23273798.2017.1381748.

Casasanto, D., and L, Boroditsky. 2008. Time in the mind: Using space to think about time. *Cognition* 106: 579-593. https://doi.org/10.1016/j.cognition.2007.03.004.

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

Crane, Tim and Craig French. 2021. The problem of perception. *The Stanford Encyclopedia of Philosophy*, ed. Edward Zalta. https://plato.stanford.edu/archives/fall2021/entries/perception-problem.

Das, D., D. Chen, A. Martins, and N. Schneider. 2014. Frame-semantic parsing. *Computational Linguistics* 40(1): 9-56. https://doi.org/10.1162/COLI_a_00163.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* Minneapolis, MN. https://doi.org/10.18653/v1/N19-1423.

Elman, J. 2009. On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive Science* 33: 547–582. https://doi.org/10.1111/j.1551-6709.2009.01023.x.

Engelhardt, Paul E. and Fernanda Ferreira. 2016. Reaching sentence and reference meaning. In *Visually Situated Language Comprehension* ed. Pia Knoeferle, Pirita Pyykkönen-Klauck and Matthew W. Crocker, 127-150. John Benjamins.

Erk, K. 2012. Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass* 6: 635-653. https://doi.org/10.1002/lnco.362.

Erk, Katrin and Aurélie Herbelot. 2021. Probabilistic constraints for meaning in context. *Proceedings of the Society for Computation in Linguistics,* 451-453. Association for Computational Linguistics. https://aclanthology.org/2021.scil-1.55.

Ervas, Francesca, Elisabetta Gola, Antonio Ledda, and Giuseppe Sergioli. 2015. Lexical ambiguity in elementary inferences: an experimental study. *Discipline Filosofiche* 22 : 149–172.

Ervas, F., Ledda, A., Ojha, A., Pierro, G.A., Indurkhya, B. 2018. Creative argumentation: When and why people commit the metaphoric fallacy. *Frontiers in Psychology*, https://doi.org/10.3389/fpsyg.2018.01815

Evans, J.S.B.T. and K.E. Stanovich. 2013. Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science* 8(3): 223–241. https://doi.org/10.1177%2F1745691612460685.

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

Ferretti, T., K. McRae, and A. Hatherell. 2001. Integrating verbs, situation schemas, and thematic role concepts. *Journal of Memory and Language* 44: 516-547. https://psycnet.apa.org/doi/10.1006/jmla.2000.2728.

Fillmore, C.J. 1982. Frame semantics. In *Linguistics in the Morning Calm*, ed. The Linguistic Society of Korea, 111-137. Seoul: Hanshin.

Firth, J.R. 1957. A synopsis of linguistic theory 1930–55. *Studies in Linguistic Analysis* 1-32. Oxford: Blackwell.

Fischer, Eugen. 2011. *Philosophical Delusion and its Therapy*. New York: Routledge.

Fischer, Eugen. 2014. Verbal fallacies and philosophical intuitions: The continuing relevance of ordinary language analysis. In *Austin on Language*, ed. Brian Garvey, 124-140. Palgrave.

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

Fischer, Eugen, and Paul E. Engelhardt. 2017a. Diagnostic experimental philosophy. *Teorema*, 36(3): 117-137.

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

Fischer, Eugen, and Paul E. Engelhardt. 2019. Eyes as windows to minds: Psycholinguistics for experimental philosophy. In *Methodological Advances in Experimental Philosophy,* ed. Eugen Fischer and Mark Curtis, 43-100. Bloomsbury.

Fischer, E., and P.E. Engelhardt. 2020. Lingering stereotypes: Salience bias in Philosophical argument. *Mind and Language* 35: 415-449. https://doi.org/10.1111/mila.12249.

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

Fischer, Eugen, Paul E. Engelhardt, Joachim Horvath and Hiroshi Ohtani. 2021. Experimental ordinary language philosophy: A cross-linguistic study of defeasible default inferences. *Synthese* 198: 1029–1070. https://doi.org/10.1007/s11229-019-02081-4.

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

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

Gentner, D., M. Imai, and L. Boroditsky. 2002. As time goes by: Evidence for two systems in processing space time metaphors. *Language and Cognitive Processes* 17: 537-565. https://doi.org/10.1080/01690960143000317.

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

Hampton, James. 2006. Concepts as prototypes. In *The Psychology of Learning and Motivation: Advances in Research and Theory,* ed. Brian Ross, 79-113. Elsevier.

Hansen, N., and E. Chemla. 2015. Linguistic experiments and ordinary language philosophy. *Ratio* 28: 422–445. https://doi.org/10.1111/rati.12112.

Hare, M., M. Jones, C. Thomson, S. Kelly, and K. McRae. 2009. Activating event knowledge. *Cognition* 111(2): 151–167. https://psycnet.apa.org/doi/10.1016/j.cognition.2009.01.009.

Hume, David. 1975 [1777]. *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*, ed. P.H. Nidditch. Oxford: Oxford University Press.

Kahneman, Daniel. 1973. *Attention and Effort*. Engelwood Cliffs, NJ: Prentice Hall.

Klepousniotou, E., B. Pike, K. Steinhauer, and V. Gracco. 2012. Not all ambiguous words are created equal: an EEG investigation of homonymy and polysemy. *Brain and Language* 12: 11-21. https://doi.org/10.1016/j.bandl.2012.06.007.

Kurtas, M., and K. Federmeier. 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. https://www.annualreviews.org/doi/10.1146/annurev.psych.093008.131123

Laeng, B., S. Sirois, and G. Gredebäck. 2012. Pupillometry: A window to the preconscious? *Perspectives on Psychological Science* 7: 18-27. https://doi.org/10.1177/1745691611427305.

Lenci, A. 2018. Distributional models of word meaning. *Annual Review of Linguistics*, 4: 151-171. https://doi.org/10.1146/annurev-linguistics-030514-125254.

Levinson, Stephen C. 2002. *Presumptive meanings. The theory of generalized conversational implicature*. Cambridge, Mass.: MIT Press.

Lewinski, M. 2012. The paradox of charity. *Informal Logic* 32: 403-439. https://doi.org/10.22329/il.v32i4.3620.

Livengood, J., and J. Sytsma. 2020. Actual causation and compositionality. *Philosophy of Science* 87: 43-69. https://doi.org/10.1086/706085.

Livengood, J., J. Sytsma, and D. Rose. 2017. Following the FAD: Folk attributions and theories of actual causation. *Review of Philosophy and Psychology* 8: 274-294. https://doi.org/10.1007/s13164-016-0316-1.

Longworth, G. 2018. The ordinary and the experimental: Cook Wilson and Austin on method in philosophy. *British Journal for the History of Philosophy* 26: 939-960. https://doi.org/10.1080/09608788.2017.1413539

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

Machery, Edouard. 2015. By default: Concepts are accessed in a context-independent manner. In *The Conceptual Mind: New Directions in the Study of Concepts,* ed. Eric Margolis and Stephen Laurence, 567-588. Cambridge, MA: MIT Press.

Machery, Edouard. 2017. *Philosophy within its Proper Bounds*. Oxford: Oxford University Press.

Macpherson, Fiona. 2013. The philosophy and psychology of hallucination. In *Hallucination: Philosophy and psychology* ed. Fiona MacPherson and Dimitris Platchias, 1-38. Cambridge, Mass.: MIT Press.

Mallon, Ron 2016. Experimental philosophy. In *Oxford Handbook of Philosophical Methodology,* ed. Herman Cappelen, Tamar Szabó Gendler, and John Hawthorne, 410-433. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199668779.001.0001

Matsuki, K., T. Chow, M. Hare, J. Elman, C. Scheepers and K. McRae. 2011. Event-based plausibility immediately influences on-line language comprehension. *Journal of Experimental Psychology: Learning, Memory and Cognition* 37: 913–934. https://psycnet.apa.org/doi/10.1037/a0022964.

McRae, K., T.R. Ferretti, and L. Amyote. 1997. Thematic roles as verb-specific concepts. *Language and Cognitive Processes* 12: 137-176. https://psycnet.apa.org/doi/10.1080/016909697386835.

McRae, K., M. Hare, J. Elman, and T.R. Ferretti. 2005. A basis for generating expectancies for verbs from nouns. *Memory & Cognition* 33: 1174-1184. https://doi.org/10.3758/BF03193221.

McRae, Ken, and Michael Jones. 2013. Semantic memory. In *Oxford Handbook of Cognitive Psychology,* ed. Daniel Reisberg, 206–219. Oxford: Oxford University Press. https://psycnet.apa.org/doi/10.1093/oxfordhb/9780195376746.013.0014.

Mehler, J., N. Sebastian, G. Altmann , E. Dupoux, A. Christophe, and C. Pallier. 1993. Understanding compressed sentences: The role of rhythm and meaning. *Annals of the New York Academy of Sciences* 682: 272–282. http://dx.doi.org/10.1111/j.1749-6632.1993.tb22975.x.

Mikolov, Tomas, Kai Chen, Greg Corrado and Jeffery Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of the International Conference on Learning Representations* (ICLR 2013), Scottsdale, Arizona. https://doi.org/10.48550/arXiv.1301.3781.

Murphy, Taylor. 2014. Experimental philosophy 1935-1965. In *Oxford Studies in Experimental Philosophy: Vol. 1*, ed. Joshua Knobe, Tania Lomboro and Shaun Nichols, 325-368. Oxford: Oxford University Press. DOI:10.1093/acprof:oso/9780198718765.001.0001

Oden, G.C., and J.L. Spira. 1983. Influence of context on the activation and selection of ambiguous word senses. *Quarterly Journal of Experimental Psychology* 35A: 51–64. https://psycnet.apa.org/doi/10.1080/14640748308402116.

Papineau, D. 2009. The poverty of analysis. *Proceedings of the Aristotelian Society* 83(1): 1-30. https://doi.org/10.1111/j.1467-8349.2009.00170.x.

Patson, N., and T. Warren. 2010. Eye movements to plausibility violations. *Quarterly Journal of Experimental Psychology* 63: 1516-1532. https://psycnet.apa.org/doi/10.1080/17470210903380822.

Peters, M.E., W. Ammar, C. Bhagavatula, and R. Power. 2017. Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. https://doi.org/10.48550/arXiv.1705.00108.

Pfeifer, N. 2012. Experiments on Aristotle's Thesis: Towards an experimental philosophy of conditionals. *The Monist* 95: 223-240. https://doi.org/10.5840/monist201295213

Pfeifer, N., and L. Tulkki. 2017. Conditionals, counterfactuals, and rational reasoning. An experimental study on basic principles. *Minds and Machines* 27: 119-165. https://doi.org/10.1007/s11023-017-9425-6

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

Rayner, K., T. Warren, B.J. Juhasz, and S.P. Liversedge. 2004. The effect of plausibility on eye movements in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 30: 1290–1301. https://psycnet.apa.org/doi/10.1037/0278-7393.30.6.1290.

Robinson, Howard. 1994. *Perception*. London: Routledge.

Ryle, Gilbert. 1949. *The Concept of Mind*. London: Hutchinsons.

Ryle, Gilbert. 1954. *Dilemmas*. Cambridge: Cambridge University Press.

Schroeder, Severin. 2006. *Wittgenstein*. Cambridge: Polity.

Sereno, S., P.J. O'Donnell, and K. Rayner. 2006. Eye movements and lexical ambiguity resolution: Investigating the subordinate bias effect. *Journal of Experimental Psychology: Human Perception and Performance 32*: 335-350. https://psycnet.apa.org/doi/10.1037/0096-1523.32.2.335.

Skovgaard-Olsen, N., D. Kellen, U. Hahn, and K.C. Klauer. 2019. Norm conflicts and conditionals. *Psychological Review 126*(5): 611–633. https://psycnet.apa.org/doi/10.1037/rev0000150

Smith, Arthur D. 2002. *The Problem of Perception*. Cambridge, Mass: Harvard UP.

Taylor, S.E., and S.T. Fiske. 1978. Salience, attention, and attribution: Top of the head phenomena. *Advances in Experimental Social Psychology*. 11: 249-288. https://doi.org/10.1016/S0065-2601(08)60009-X.

Thagard, P., and R.E. Nisbett. 1983. Rationality and charity. *Philosophy of Science* 50: 250-267. https://doi.org/10.1086/289108.

Urmson, James O. 1969. A symposium on Austin's method. Part I. In *Symposium on J.L. Austin,* ed. K.T. Fann. 76-86. Routledge.

Waismann, Friedrich. 1968. *Principles of Linguistic Philosophy*. ed. Rom Harré. London: St. Martin's Press.

Wheeldon, L.R., and W.J.M. Levelt. 1995. Monitoring the time course of phonological encoding. *Journal of Memory and Language* 34(3): 311-334. https://psycnet.apa.org/doi/10.1006/jmla.1995.1014.

Wittgenstein, Ludwig. 1953. *Philosophical Investigations*. Oxford: Blackwell.

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