Synthese
https://doi.org/10.1007/s11229-020-02708-x

Inappropriate stereotypical inferences? An adversarial collaboration in experimental ordinary language philosophy

Eugen Fischer[1] · **Paul E. Engelhardt**[2] · **Justin Sytsma**[3]

Received: 25 October 2019 / Accepted: 16 May 2020

Abstract
This paper trials new experimental methods for the analysis of natural language reasoning and the (re)development of critical ordinary language philosophy in the wake of J.L. Austin. Philosophical arguments and thought experiments are strongly shaped by default pragmatic inferences, including stereotypical inferences. Austin suggested that contextually inappropriate stereotypical inferences are at the root of some philosophical paradoxes and problems, and that these can be resolved by exposing those verbal fallacies. This paper builds on recent efforts to empirically document inappropriate stereotypical inferences that may drive philosophical arguments. We demonstrate that previously employed questionnaire-based output measures do not suffice to exclude relevant confounds. We then report an experiment that combines reading time measurements with plausibility ratings. The study seeks to provide evidence of inappropriate stereotypical inferences from appearance verbs that have been suggested to lie at the root of the influential 'argument from illusion'. Our findings support a diagnostic reconstruction of this argument. They provide the missing component for proof of concept for an experimental implementation of critical ordinary language philosophy that is in line with the ambitions of current 'evidential' experimental philosophy.

Keywords Experimental philosophy · Ordinary language philosophy · Reading time measurements · Stereotypical inference · Appearance verbs · Argument from illusion

✉ Eugen Fischer
e.fischer@uea.ac.uk

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

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

3 Department of Philosophy, Victoria University of Wellington, PO Box 600, Wellington 6140, New Zealand

Published online: 03 June 2020

1 Introduction

Natural language reasoning is shaped by inferences that have received relatively little attention in the philosophical literature: Arguments couched in natural language are strongly shaped by automatic inferences that continuously occur in language comprehension and production. Default pragmatic inferences, including stereotypical inferences (Levinson 2000), enrich our spontaneous understanding of verbal case descriptions and premises of arguments. These defeasible default inferences shape philosophical thought experiments (Saint-Germier 2019) and arguments couched in natural language (Fischer and Engelhardt 2019a). J. L. Austin (1962) mooted the idea that they sometimes do so not for better but for worse. He suggested that philosophers sometimes make defeasible stereotypical inferences even in contexts that defeat them; that such contextually inappropriate inferences are particularly prone to occur when we give special uses to words that have related but distinct uses in ordinary discourse; and that such fallacies are at the root of some influential philosophical paradoxes and problems. The approach of 'critical' ordinary language philosophy seeks to resolve pertinent problems by exposing fallacies in the underlying reasoning. Its first paradigm (worked model) was Austin's *Sense and Sensibilia* (1962), which sought to contribute to 'dissolving' what is today known as 'the problem of perception' (Crane and French 2015; Smith 2002) by exposing fallacies in the underlying 'argument from illusion'.

The present paper will use this paradigm to explain the approach (Sect. 1.1), and why it needs empirical vindication (Sect. 1.2). The bulk of the paper will then argue that such vindication requires sophisticated experimental methods from psycholinguistics and will employ reading-time measurements with eye tracking for the purpose. The philosophical upshot will be the vindication of a diagnostic reconstruction of the argument from illusion that exposes contextually inappropriate stereotypical inferences at its root. Methodologically, the paper will show how eye tracking can be employed for the new, philosophical, purpose of argument analysis. At the metaphilosophical level, the paper will thus afford proof of concept for critical ordinary language philosophy as a partially experimental enterprise.

1.1 The target argument

The argument from illusion is the historically most influential argument against naïve realism in the philosophy of perception (Robinson 1994). The argument proceeds in two main steps: Its 'base step' argues that in particular cases of non-veridical perception ('illusion'), where something appears F but actually has the different colour, size, or shape G, subjects are (directly) aware of sense-data, rather than physical objects; a 'spreading step' extends this conclusion to all cases of perception (Snowdon 1992). Along with parallel arguments (e.g., 'from hallucination'), the argument thus appears to challenge our common-sense conception of perception, according to which our senses typically provide us with unmediated awareness of physical objects. This challenge gives rise to the philosophical 'problem of perception' (Crane and French 2015; Smith 2002).

Austin (1962) engaged with early twentieth century proponents of the argument (incl. Ayer 1940, pp. 3–11; Broad 1923, p. 240; Price 1932, pp. 27–30; Russell 1912, pp. 1–3). These thinkers leaped from initial case descriptions to negative conclusions, e.g.:

(1) When a subject looks at a round coin sideways, the coin appears elliptical to her.
(2) When a subject looks at a round coin sideways, she is not (directly) aware of the round coin.

They then inferred that subjects are, instead, aware of an 'elliptical sense-datum', from an uncontroversial response to (2):

(3) When a subject looks at a round coin sideways, she is (directly) aware of *something*.
(4) By (2) and (3), the subject is then (directly) aware of something other than the round coin (namely, a 'sense-datum').

Later, we will consider the currently canonical form of the base step and see that it, too, ultimately relies on the spontaneous inference from the initial case description (1) to the negative conclusion (2) explicitly acknowledged by the early twentieth century version of the argument.

Austin seems to suggest that this key move involves contextually inappropriate stereotypical inferences from the appearance verb used in (1)—mainly 'appear' (e.g., Ayer 1940, p. 3; Fish 2010, pp. 12–13; Robinson 1994, p. 57; Russell 1912, p. 2; Smith 2002, p. 25) or 'seem' (e.g., Ayer 1940, p. 3; Broad 1923, pp. 239–240; Crane and French 2015, p. 3; Moore 1918/19, pp. 21–23; Russell 1912, p. 2), and only occasionally 'look' (e.g., Ayer 1940, p. 4). Austin (1962, pp. 36–37) suggests that while 'look' is typically used simply to comment on the look of things, 'appear' 'would typically be used with reference to certain *special circumstances*' affecting judgment, and 'seem' 'makes an implicit reference to certain [inconclusive] *evidence*' supporting judgment. Accordingly, at any rate 'X appears F [to S]' and 'X seems F [to S]' are typically used with doxastic implications (S thinks that X is F) (*cf.* Brogaard 2013, 2014) and have us infer from (1) above that

(C) The viewer thinks that the object viewed is elliptical.

In the context of the argument from illusion, such doxastic inferences would be doubly inappropriate. First, proponents of the argument explicitly acknowledge that, in the familiar cases at issue, viewers confidently judge that things actually have some shape, size, or colour distinct from the one they look under the circumstances (e.g., Ayer 1956, p. 88; Broad 1923, pp. 236–237, 241; *cf.* Price 1932, p. 27); nobody falls prey to an 'illusion', in these cases. Second, proponents intend to use appearance- and perception-verbs in a rare 'phenomenal' sense that implies nothing about perceivers' physical environment or their knowledge or beliefs about it (Ayer 1956, p. 90; Jackson 1977, pp. 33–49; *cf.* Chisholm 1957, pp. 44–48; Maund 1986).[1] This sense—and only

[1] We provide experimental evidence that a suitably non-factive, non-epistemic, and non-doxastic sense (where 'X looks F to S' etc. imply neither that X is F nor that S knows or thinks it is F) is recognised for all three appearance verbs not only by (some) philosophers but by ordinary speakers (see Appendix 3). We will refer to this rare, but established sense as 'phenomenal'. For this paper's purposes, it is immaterial how well the cited philosophical explanations capture—or how much they misconstrue—it.

this sense—allows them to make claim (1). But this sense does not license the inference from (1) to (C).

Austin does not explain how this doxastic inference could contribute to the argument's move from (1) to (2). Fischer (2014) proposes to fill the gap by showing how (C) and uncontroversial background assumptions entail (2): (1) entails that the coin viewed is round. Together with (C), this entails that the viewer has a wrong belief about the coin, and does not know that it is round, or that there is a round coin. Together with standard definitions of 'to be aware of' ('to have cognizance, know, have knowledge as obtained by observation or information', *Oxford English Dictionary*), this entails that the viewer is not aware of the round coin. Standard definitions of 'direct awareness' do not cancel epistemic implications but rather impose the stricter requirement that the relevant knowledge be acquired without conscious inference. Hence the ignorant viewer is not 'directly aware' of the round coin, either. If this a priori reconstruction of the move from (1) to (2) is correct, the argument from illusion proceeds from a contextually inappropriate inference from the appearance verb in its initial premise[2]—so that the argument does not get off the ground and the philosophical problem of perception fails to arise, if parallel arguments similarly fail.

1.2 Need for empirical vindication

As developed by Austin, critical ordinary language philosophy seeks to expose 'seductive (mainly verbal) fallacies' as 'concealed motives' for formulating philosophical paradoxes and problems that can be 'dissolved' through such exposure (Austin 1962, p. 5).[3] That the inferences are 'concealed' means that thinkers are not conscious of making them and presupposing their conclusions, in the relevant arguments.[4] Psycholinguistic research has confirmed the occurrence of pertinent inferences: automatic inferences from words that continually occur beyond conscious awareness in language comprehension and production (see below, Sect. 2.1), which are crucially involved in verbal reasoning. This opens the intriguing possibility that philosophers might inadvertently but recurrently rely on fallacious automatic inferences they would explicitly reject.

Any hypothesis that such inferences drive a philosophical argument requires empirical vindication. The need for empirical vindication arises from the facts that the hypothesized inferences are, first, 'concealed' and, second, fallacious. First, an a priori reconstruction of verbal reasoning (like the one above) can only specify one of several inference chains that *could* have led thinkers from a premise (like 1) to a conclusion it does not entail (like 2). Thinkers have no privileged access to automatic inferences. Where thinkers' relevant inferences (e.g., from 1 to C) are automatic, thinkers' self-reports or acceptance of a proposed a priori reconstruction therefore

[2] Since proponents of the argument intend to use perception-verbs, too, in a phenomenal sense, the argument also involves undue reliance on the ordinary epistemic sense of 'to be aware of'. This vitiates those rare versions of the argument that proceed from unfamiliar cases of non-veridical perception.

[3] For a variety of different recent contributions to 'critical' ordinary language philosophy, see Baz (2017), Fischer et al. (2019), and Hansen (2018), *cf.* Weinberg (2017).

[4] This is the most productive and charitable interpretation, in that it does not take Austin's project to rest on the accusation that the targeted philosophers dishonestly 'conceal' their motives from others.

cannot provide a justified answer to the question of which inference chain—of those many potentially relevant chains—actually led them from premise to conclusion.[5] Automatic inferences can, however, be documented experimentally, so we can turn to experiments for a justified answer.

Second, and independently, principles of charity limit the extent to which readers of philosophical texts may attribute fallacious inferences to authors. Plausible 'medium-strength' principles strike a balance between respecting authors' rationality and appreciating human fallibility: They allow us to attribute violations of rational or linguistic norms to competent thinkers and speakers, but only if we have an empirically supported explanation of when and why competent thinkers commit the fallacies at issue (Thagard and Nisbett 1983). Austin seeks to expose contextually inappropriate stereotypical inferences in some influential philosophical arguments. These fallacies are a case in point: Austin maintains—and psycholinguistic research confirms (see Sect. 2.1)—that competent speakers' inferences are highly sensitive to contextual cues. But philosophers are competent speakers. The *prima facie* uncharitable hypothesis that a highly influential philosophical argument relies on a contextually inappropriate stereotypical inference is hence in need of empirical support.

The required support is twofold: First, we need to develop an empirically supported explanation of when (under what conditions) and why competent speakers make inappropriate stereotypical inferences. The empirical support for this explanation will include experimental findings that document contextually inappropriate stereotypical inferences which occur under the conditions C where the explanation predicts them. Based on textual evidence and a priori logical reconstruction, a diagnostic reconstruction may suggest a particular argument involves a specific such fallacy. To support this suggestion, we then need, in addition, to show that the relevant conditions C prevail in the formulation of the argument, and provide experimental evidence that the specific inferences posited in it (e.g., doxastic inferences from phenomenal uses of appearance verbs) are actually made by competent speakers/thinkers.

In a novel extension of the 'evidential program' in experimental philosophy (Sytsma and Livengood 2016, pp. 40–42),[6] some recent studies have sought to provide such empirical foundations for a 'critical' ordinary language philosophy (Fischer and Engelhardt 2017a; Fischer et al. 2019). They developed a psycholinguistic explanation that lets us understand when and why competent speakers (like philosophers) cannot help making contextually inappropriate stereotypical inferences from words with distinct, but related senses (Fischer and Engelhardt 2019a, b). They used advanced psycholinguistic methods (pupillometry and reading time measurements with eye tracking) to support this explanation and document contextually inappropriate stereotypical inferences predicted by it. However, while they suggested that similar inferences could occur in philosophical arguments, so far only two questionnaire-based studies (Fis-

[5] Acceptance of such a reconstruction can amount to adoption of a new, more fully spelled out argument, but cannot establish that a particular automatic inference was involved in the original argument.

[6] To date, evidential experimental philosophy focused on assessing the evidentiary value of intuitive judgments about verbally described cases, mainly in order to evaluate thought experiments employing the 'method of cases' (Machery 2017). Some proponents of the research program have advocated its extension to the assessment of premises (Nado 2016) and inferences (Fischer and Engelhardt 2019a) in philosophical arguments beyond thought experiments.

cher and Engelhardt 2016; Fischer et al. 2019) attempted to provide evidence of the specific inferences that had been hypothesised to occur in an influential philosophical argument.

1.3 Aim and agenda

The present paper re-examines the hypothesis that competent language users make doxastic inferences from phenomenal uses of appearance verbs—like the inference from (1) to (C) above—*which are cancelled by the context* but influence further judgment and reasoning anyway. We report an adversarial collaboration that first exposes a gap in the extant empirical argument and then closes the gap. We will argue that extant questionnaire-based studies are potentially subject to a confound. To exclude pertinent confounds, we suggest studying automatic inferences by combining different questionnaire-based outcome ('offline') measures with each other and with process ('online') measures that can tap into cognitive processes as they unfold. The main study will show how online eye-tracking methods from psycholinguistics, which are commonly used to examine general hypotheses about language processing, can be adapted to study how people reason with or about specific lexical items of philosophical interest.

This is the first study to use this methodology to document automatic comprehension inferences that drive philosophical arguments. The paper will thus present new tools for the analysis of natural language reasoning in philosophy, in line with ongoing efforts to add new empirical methods and fresh philosophical aims to the repertoire of experimental philosophy (review: Fischer and Curtis 2019). Their application will provide the first compelling evidence of contextually inappropriate stereotypical inferences that drive an influential philosophical argument. We thus seek to provide what is still missing for successful proof of concept for experimental implementation of ordinary language philosophy's critical project.

Section 2 will explain the notion of 'stereotypical inference', set out conditions under which contextually inappropriate stereotypical inferences occur, and summarise previous experimental work that sought to document inappropriate stereotypical inferences from appearance verbs. Section 3 will develop an alternative explanation of the previous experimental findings and will report three new experiments that provide initial support for this alternative explanation. To adjudicate between these two explanations, Sect. 4 will report an eye tracking study and examine correlations between different offline measures. Section 5 will discuss the intended philosophical application of the findings: how they help us expose pernicious inferences in the argument from illusion.

2 Inappropriate stereotypical inferences?

2.1 Stereotypical inferences

Stereotypes are implicit knowledge structures in semantic memory that are built up from observation of co-occurrence frequencies in the physical and discourse environment (In the supermarket, most tomatoes are red and round) (McRae and Jones 2013). They may be associated with specific words and are also known as 'prototypes' and 'situation schemas' when associated with (object-) nouns and verbs, respectively. When we hear words, the most stereotypical features come to mind first and are easiest to process; these features are diagnostic or predictive of the relevant categories (Hampton 2006). Event nouns (Hare et al. 2009) and verbs (Ferretti et al. 2001) are associated with complex *situation schemas* which include typical features of events or actions (instruments used, etc.), agents, and 'patients' acted on. Priming studies have shown that single words ('tomato') activate stereotypical features (*red*) rapidly (within 250 ms) (review: Engelhardt and Ferreira 2016). In sentence comprehension, feature activation depends upon thematic fit: Sentence fragments that leave blank the agent- and the patient-role, respectively, activate typical features of agents and patients, respectively ('She was arrested by the ___' activates *cop*, not *crook*) (Ferretti et al. 2001; *cf.* Kim et al. 2016).

In utterance interpretation, the implicit knowledge encoded by stereotypes is immediately brought to bear to fill in details that remain unstated (e.g., the tomato referred to will have been red, the secretary female, etc.) (Levinson 2000). Stereotypical associations support defeasible spontaneous inferences from words ('tomato', 'secretary') to features stereotypically associated with them, as illustrated by common responses to the following vignette (from Giora 2003, p. 3):

> A young man and his father had a severe car accident. The father died, and the young man was rushed to hospital. The surgeon at the emergency room refused to operate on him, saying, 'I can't. He's my son.' – How is this possible?

When first encountering the vignette, many readers have difficulty answering this question. This difficulty is due to an automatic inference from 'surgeon' to the stereotypically associated gender, whose conclusion (The surgeon is male) may initially get presupposed in further reasoning (therefore, the surgeon is the young man's father), despite the resulting inconsistency with the context ('the father died').

To study such inferences, psycholinguists 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. If the hypothesised inference is made, the clash of the conclusion with the sequel will engender comprehension difficulties requiring cognitive effort. This effort is picked up by a variety of process measures including pupil dilations (Sirois and Brisson 2014), longer reading times (Clifton et al. 2007), and signature electrophysiological responses ('N400s') (Kutas and Federmeier 2011). Studies using priming or the cancellation paradigm have shown that verbs prompt parallel probabilistic inferences to typical features of agents, patients, and instruments (Ferretti et al. 2001; Harmon-Vukic et al. 2009; Welke et al. 2015).

In addition to schemas associated with individual verbs, we have schemas that 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), even in the absence of single-word priming of typical patients (Matsuki et al. 2011). That is, the words we hear or read activate 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. 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 incremental utterance interpretation, ever more specific schemas are thus activated by verbs in conjunction with subject- and object-nouns, with prepositions and syntactic constructions like verb aspect (Ferretti et al. 2007), and with simultaneous visual stimuli (Kamide et al. 2003).

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; *cf.* Garrett and Harnish 2007): Speakers typically skip mention of stereotypical features but make deviations from stereotypes explicit. In the absence of such explicit indications to the contrary, hearers assume the situation talked about conforms to the relevant schemas, deploy the most specific schemas relevant, and fill in detail in line with this knowledge about situations of the kind at issue. The rapid deployment of the most specific schemas together with the explicit marking of stereotype-deviations ensures that stereotypical enrichment is highly context-sensitive and inappropriate stereotypical inferences hinder comprehension and further reasoning only rarely.

This is good news for language users but *prima facie* bad news for the critical project of ordinary language philosophy: It means that in the absence of specific empirical reasons we have no right to believe that competent speakers like philosophers should rely on contextually inappropriate stereotypical inferences in their arguments.

2.2 Contextual impropriety?

Against this backdrop, recent work in experimental philosophy has identified specific conditions under which competent language users make inappropriate stereotypical inferences that stubbornly influence further cognition. Under these conditions, language users do not merely make contextually inappropriate stereotypical inferences, but are unable to disregard their conclusions even after the problematic inferences have been made explicit. In this way these stubborn inferences contrast, for instance, with the problematic inference from 'surgeon' in the vignette above, where readers have no problem disregarding the problematic stereotypical inference, once it has been made explicit.

The stubborn inferences at issue are from words with distinct but related senses (polysemes). Whereas words with different unrelated meanings (homonyms) activate separately represented and mutually exclusive stereotypes that compete for sustained activation (Beretta et al. 2005; Pylkkänen et al. 2006), polysemes typically activate

a unified 'core representation' (Klepousniotou et al. 2012; MacGregor et al. 2015). Often, this is the stereotype associated with the dominant sense (or a sense privileged, e.g., through embodiment), which is then deployed to interpret utterances that use the word in a less salient sense. According to the *Graded Salience Hypothesis*, the stereotype associated with the most '*salient*', i.e., most frequently encountered and prototypical,[7] sense is activated most swiftly and strongly by the verbal stimulus, irrespective of context (Fein et al. 2015; Giora 2003). According to the *Retention/Suppression Hypothesis*, readers/hearers often interpret utterances that employ the word in a less frequent sense by retaining that initially most strongly activated stereotype and suppressing its contextually irrelevant components (Giora 2003; Giora et al. 2014). To interpret, for example, the metaphorical epistemic use of 'see' in 'I see your point', hearers retain the situation schema associated with the dominant visual use of the word, with agent-features including *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* and *X is near S,* and then attempt to suppress all component features of the schema, except the epistemic agent features.

But suppression of irrelevant components may remain partial: Frequently co-instantiated core components of the stereotype laterally pass on activation among each other (Hare et al. 2009; McRae et al. 2005). Where some, but not all of them are contextually relevant, such lateral cross-activation of irrelevant components will complement their initial strong activation due to salience and render their complete suppression impossible. When they are only partially suppressed, schema components continue to support stereotypical inferences. The psycholinguistic research reviewed has motivated the *Salience Bias Hypothesis* (Fischer and Engelhardt 2019a, b): When

(i) one sense of a polysemous 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 core 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.

When stereotypical inferences that are initially triggered—as per (1)—clash with contextual information or background beliefs, they can be suppressed within one second and before they influence further judgment and reasoning (Fischer and Engelhardt 2017b). (2) hypothesises that this does not happen, where conditions (i)–(iii) are met.

Where less salient uses are associated with distinct stereotypes, explicit marking of the less salient use (through riders like 'figuratively speaking', 'in a special sense', etc.) reinforces the activation of relevant stereotypes and helps them win the competition for activation against dominant stereotypes that initially receive stronger activation from

[7] A sense of a polysemous word (e.g., 'see') is more or less 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*). This is typically assessed with a sentence completion task (Chang 1986).

the verbal stimulus (Givoni et al. 2013). Sometimes, such reinforcement can thus prevent inappropriate inferences. To prevent the inferences posited by the Salience Bias Hypothesis, however, explicit marking would need to reinforce suppression of components of the dominant schema, rather than activation of competitors. Therefore, these inferences cannot be prevented by explicit marking of the less salient use.

The first experiments to examine the Salience Bias Hypothesis provided supporting evidence from perception verbs, and illustrate the interplay between these conditions: An eye-tracking study revealed extensive similarities in intricate processing patterns for 'aware'- and 'see'-sentences that strongly suggest similar schemas are deployed in interpreting them (Fischer and Engelhardt 2019b) with the Retention/Suppression strategy [as per condition (ii)]. However, a corpus analysis on random 1000-sentence samples from the *British National Corpus* confirmed that while the visual use is clearly the most frequent for 'S sees X' (68% of sampled occurrences) [as per (i)], it is far less frequent for 'S is aware of X' (23%) (Fischer and Engelhardt 2017a).[8] Three studies that combined plausibility ratings with either pupillometry or reading time measurements then provided evidence that competent speakers make contextually inappropriate inferences from purely epistemic uses of 'S sees X' to spatial conclusions (*X is in front of S*) unless the contextual relevance of typically co-occurring core schema components is minimised [*cf.* (iii)], but make such inappropriate inferences from purely epistemic uses of 'S is aware of X' [whose visual use is not dominant, *cf*. (i)] only where the contextual relevance of such core schema components is maximised (Fischer and Engelhardt 2017b, 2019a, b). These inferences influenced further judgment, even though participants drawn from the same population evinced explicit knowledge that such inferences typically fail to lead from true premises to true conclusions (Fischer and Engelhardt 2019a, pre-study).

The Salience Bias Hypothesis identifies a first set of conditions under which competent speakers cannot help going along with stereotypical inferences they know to be inappropriate. It thus provides first empirical foundations for the approach of critical ordinary language philosophy that seeks to expose in philosophical arguments 'seductive (mainly verbal) fallacies' that can be operative as 'concealed motives' (Austin 1962, p. 5). To apply the approach to our target, the argument from illusion, and vindicate the proposed diagnostic reconstruction (Sect. 1.1), we need to show that these conditions apply to appearance verbs, and experimentally document inappropriate doxastic inferences from phenomenal uses of these verbs, such as those in the argument's initial premise.

2.3 Inferences from appearance verbs: hypothesis and previous experiment

In their philosophically relevant intransitive sense ('Joe looks dirty'), appearance verbs function as subject-raising verbs that are semantically unrelated to their grammatical subjects ('Joe') and serve not so much to predicate any property from their complement (*dirtiness*) of those subjects' referents (Joe) as to attribute to the often implicit patient an

[8] Similarly, the visual use is clearly the most prototypical for 'see', but less pronounced for 'is aware', as evidenced by 94% versus 46% pertinent completions in a sentence completion task (Fischer and Engelhardt 2017a).

experiential, epistemic, or doxastic attitude towards a content (*Joe is dirty*) (Brogaard 2013, 2014). A distributional semantic analysis of the words' intransitive use in a parsed Wikipedia snapshot suggests that 'seem' and 'appear', and to a lesser extent 'look', are most frequently used to attribute doxastic attitudes, less frequently used to attribute epistemic attitudes, and yet less frequently to attribute experiential attitudes (Fischer et al. 2015). In their intransitive use, all three verbs share the same sense ('give a certain impression or have a certain outward aspect', *WordNet* 3.1), which is far more frequent than any other sense associated with an intransitive use (ibid.). This suggests that, in conjunction with the relevant syntactic cues (Goldberg 2003), all three verbs rapidly activate the same associated situation schema, or very similar schemas. It further suggests that doxastic, epistemic, and experiential patient features are integrated with decreasing strength into this 'appearance schema', or are integrated in this order of strength, but with slightly different weightings, into similar schemas.

Fischer and colleagues (2019) hypothesise that this dominant schema is deployed to interpret the phenomenal use of appearance-verbs, with the Retention/Suppression strategy: This strategy retains the experiential component of the appearance schema (*S looks at X, X visually looks F to S*), and attempts to suppress the remaining core schema components (*S thinks X is F, S knows X is F*). The Salience Bias Hypothesis then predicts that

H_1 Phenomenal uses of appearance verbs ('X seems F [to S]', etc.) will [1] trigger doxastic inferences (to *S thinks that X is F*) that [2] influence further judgment and reasoning, even in contexts in which these inferences are explicitly cancelled.

To experimentally examine this hypothesis, Fischer et al. (2019) implemented the cancellation paradigm (Sect. 2.1) with a forced-choice plausibility ranking task. In their study, a questionnaire presents participants with short texts that differ in one critical word, such as:

6a The hill seemed quite steep. The rambler thought it was gentle.
6b The hill was quite steep. The rambler thought it was gentle.

Participants then judge which of the two strike them as more plausible. Critical items pair sentences using 'look', 'appear', or 'seem' with otherwise identical sentences that employ the contrast verb 'is' which lacks doxastic implications. 'Is'-texts are mildly implausible, insofar as they claim that the viewer got quite obvious things wrong. If appearance verbs are understood to take the subsequently mentioned protagonist as a patient (*The hill seemed steep to the rambler*) and—as per H_1—trigger doxastic inferences (*The rambler thought the hill was steep*) that remain unsuppressed and influence further cognition, the persistent clash with the sequel is felt to engender a contradiction. Participants then judge mildly implausible 'is'-texts (like 6b) more plausible than outright contradictory appearance-sentences (like 6a). In this experiment, the conflict with the sequel renders a phenomenal reinterpretation of the appearance verb contextually appropriate. On such reinterpretation, the appearance sentence describes only the protagonist's (the rambler's) experience but attributes no belief to him. This would remove the impression of a contradiction. Consistent preferences of 'is'-sentences over alternatives therefore provide first evidence that doxastic inferences are made from phenomenal uses of appearance verbs and are then maintained long enough to

influence further judgment, even in contexts in which those inferences are explicitly cancelled.

This study used the plausibility-ranking task to simultaneously test the further hypothesis that these contextually inappropriate stereotypical inferences are not defeated by competing pragmatic inferences. Participants who keenly feel the conflict between doxastic inferences and inconsistent sequels can avoid a contradictory interpretation by reassigning the patient role of the appearance verb, from the text's protagonist (e.g., the rambler) to its author (the hill seems steep, not to the rambler, but to the author of the questionnaire). This turns the first sentence into the expression of an authorial self-attribution of a belief (*I think the hill is steep*). In contrast with belief attributions to others, appearance sentences that imply such self-attributions are often used to express hedged judgments about the agent-role filler, which the author could have expressed more simply by writing, e.g., 'The hill is steep'. From preference of an appearance-verb over the simpler 'is' hearers-readers therefore infer with the Maxim of Manner that doubt-and-denial conditions obtain (it is in doubt or contention whether the hill is steep) (Grice 1961). These conditions make it more plausible that the protagonist (the rambler) should have a different belief. In this way, pragmatic inferences that are higher in the pragmatic pecking order may defeat the stereotypical inferences of interest (*cf.* Levinson 2000, pp. 157–158). By requiring comparisons of appearance and 'is'-sentences, the plausibility-ranking task invites competing Manner-inferences that would attenuate preferences for 'is'-sentences.

Fischer and colleagues assumed [a] that the patient-role reassignment that facilitates Manner inferences is more likely to occur the more contradictory an item seems to participants. They further assumed [b] that items with abstract objects ('The plan looked good. Cole believed it was terrible') would be perceived as more contradictory than items with visual objects (like 6a). They therefore added an equal number of critical items with abstract objects and predicted attenuated preferences for 'is'-sentences in these items. The attenuation of 'is'-preferences concerning items with abstract objects would show that the plausibility ranking task had managed to create conditions inviting Manner inferences, so that consistent 'is'-preferences concerning items with visual objects would show that the doxastic inferences of philosophical interest go through undefeated even under such conditions.

Fischer and colleagues examined preferences in English and in two languages with increasingly rigid verb-final sentence structure, German and Japanese, where inferences from the verb play a less central role in utterance interpretation and exert less influence on further judgment and reasoning. Their findings were consistent with their predictions: In items with visual objects (like 6), they observed consistent preferences for 'is'-sentences over counterparts with 'look', 'appear', and 'seem', and attenuated preferences in items with abstract objects, across all three languages. They interpreted these findings as evidence that the contextually inappropriate stereotypical inferences posited by H_1 are not defeated in the perceptual cases of interest and took their findings to support the initially a priori reconstruction of the 'argument from illusion' that takes its opening move to rely on such inappropriate doxastic inferences from phenomenal uses of appearance verbs (Sect. 1.1). But does this interpretation of the findings stand up to scrutiny?

3 Exploring an alternative explanation

3.1 An alternative explanation

Justin Sytsma (2019) proposed an alternative interpretation of these results. He questions the assumption that Fischer et al.'s participants base their judgments about items with visual objects on interpretations that treat the protagonist of the second sentence (e.g., the rambler) as the patient of the appearance verb in the first sentence (*The hill seemed quite steep to the rambler*). Instead, participants could treat the author as the patient. This could reflect participants' initial assignment of the patient role or its reassignment in line with Fischer et al.'s reasoning for items with abstract objects. That is, this might happen because conflicts of initial doxastic inferences (*The rambler thought the hill was quite steep)* with the sequel ('The rambler thought it was gentle') lead participants to reassign the patient role from the protagonist to the author. Accepting Fischer et al.'s reasoning, such reassignment would occur if, *contra* assumption [b] above, participants perceive the items with visual objects as no less contradictory than the items with abstract objects. Participants would thus come to interpret the appearance sentence as expressing a belief or hedged judgment of the author of the questionnaire (*I think the hill is quite steep*). If participants assign the patient role to the author, either from the start or through reassignment, the observed preferences for 'is'-sentences over appearance sentences could not be due to a persistent clash of doxastic inferences with the sequel—which attributes a belief to someone else (the protagonist).

This reasoning motivates an alternative explanation of the observed preferences: These are based on participants' assessment of the first sentence only (e.g., 'The hill was quite steep' vs. 'The hill seemed quite steep') and are driven by subjectivity judgements. 'Is'-sentences are read as stating an authorial claim about how things are (*The hill was quite steep*). In line with treating the author as the patient of the appearance verb, appearance sentences are read as expressing an authorial opinion about how things are (*I think the hill was quite steep*). When forced to assess the relative plausibility of the two, participants will therefore take into account whether the author should be making a factual claim or express his opinion about the matter at hand.[9] In the absence of discourse context, this question will plausibly be decided by objectivity/subjectivity judgments concerning the claim under discussion: Is this more objective (more a matter of fact) or more subjective (more a matter of opinion)? Whether visual objects have a certain shape, size, colour, or other visual property presumably strikes participants as more objective, leading them to prefer 'is'-sentences over appearance counterparts. Whether abstract objects have the properties the items ascribe to them (e.g., whether a plan is good) will often strike participants as more subjective, leading to attenuated preferences for 'is'-sentences.

[9] Implementations of the cancellation paradigm with plausibility assessment tasks assume that participants interpret 'plausible' as 'likely to be true'. This alternative account assumes participants will read it as 'plausible or appropriate thing to say'. In line with common psycholinguistic practice, Fischer et al. (2019) did not explain the intended meaning of 'plausible' to their participants, so cannot exclude this alternative interpretation.

On this account, the use of appearance verbs in Fischer et al.'s items are interpreted throughout as doxastic, not phenomenal: The first sentences are interpreted throughout as expressing an opinion of the author, and this doxastic attribution is not cancelled by the sequel (which attributes an opinion to another person). Therefore, the observed preferences provide no evidence that phenomenal uses of appearance verbs trigger doxastic inferences that influence further judgment and reasoning, even in contexts in which these inferences are explicitly cancelled. On this account of Fischer et al.'s findings, they fail to provide support for hypothesis H_1.

This criticism involves three hypotheses about how the items used by Fischer and colleagues (2019) are processed and assessed. Initial motivation was provided by (h1):

(h1) Even if participants initially assign the patient role of the appearance verb to the text's protagonist, the appearance items with visual objects do not strike participants as notably less contradictory than the appearance sentences with abstract objects.

In response to such conflicts, participants would reassign the patient role. Either as a result of such reassignment or from the start, the criticism holds:

(h2) Participants will tend to assign the patient-role of the appearance verb to the text's author (rather than its protagonist), across appearance items.

This would rule out Fischer et al.'s interpretation of their findings. Sytsma's alternative account of these findings then relies on a final hypothesis:

(h3) Claims about the visual objects will be deemed less subjective than claims about the abstract objects.

We conducted three experiments to examine these hypotheses.

3.2 Experiment 1

To assess h1, the first study elicited contradictoriness ratings.

106 participants were recruited through advertising on Google for a free personality test, which was administered after the main task. Participants were restricted to native English-speakers, 16 years of age or older, who passed an attention check.[10]

Participants rated slight variations of the appearance sentences from the 36 critical items used in Fischer et al. (2019). The first sentence of each text was modified to make the patient explicit, e.g.,

The hill seemed quite steep to the rambler. The rambler thought it was gentle.

Participants were instructed that both sentences in each text are about the same person and rated whether there is a contradiction between the two sentences on a scale from 1 ('no contradiction') to 7 ('complete contradiction'). As an attention check, an additional item asked participants to select '5' on the scale. Items were presented in random order.

[10] Ads were targeted to North America. Participants were 70.8% women (two non-binary), average age 46.3, ranging from 16 to 80.

The mean contradictoriness rating for each item was numerically above the neutral mid-point, with 33 of the 36 items being significantly above this point.[11] Mean ratings varied from a low of 4.03 ('Their efforts seemed idealistic to Sam. Sam thought they were self-serving.') to a high of 5.93 ('The estimate looked accurate to Anna. Anna thought it was completely wrong.') Further, for 35 of the 36 items, a majority of participants gave a response above the neutral mid-point. Thus, in line with h1, we find that participants generally treated the items as being contradictory.

Averaging across the visual items, we found a mean of 5.16, which was significantly above the neutral mid-point $t(105) = 7.7198$, $p < 0.0001$. Averaging across the abstract items, we found a mean of 5.39, which was also significantly above the neutral mid-point $t(105) = 10.023$, $p < 0.0001$.[12] While the mean for the abstract items was significantly greater than for the visual items, the effect size was negligible $t(105) = 3.8956$, $p = 0.00017$, Cohen's $d = 0.16$. Breaking the results down by appearance verb, we found this effect was driven by 'looks': For this verb, we found a significant difference with a small effect size, with a mean of 5.00 for the visual items and 5.57 for the abstract items $t(105) = 5.6647$, $p < 0.0001$, Cohen's $d = 0.36$. By contrast, no significant difference was found for either 'appears' (mean of 5.33 for visual items vs. 5.32 for abstract items $t(105) = 0.13684$, $p = 0.89$) or 'seems' (mean of 5.14 for visual items vs. 5.28 for abstract items $t(105) = 1.6181$, $p = 0.11$).

While we found that overall participants regarded abstract items as *slightly* more contradictory than visual items, the effect size was negligible and there was a significant difference for only one of three appearance verbs. These results are consistent with h1 and motivate the hypothesis that, even if participants in the study of Fischer and colleagues (2019) initially assigned the patient role of the appearance verb to the protagonist in the text, they reassigned it to the author, in appearance sentences with both abstract and with visual objects. Experiment 2 tested more directly whether participants assign the patient role to the author in both types of sentences.

3.3 Experiment 2

To assess h2, we elicited explicit patient role assignments to the appearance verbs in the critical items at issue.

We recruited 44 participants with the same approach and from the same population as in the previous experiment.[13] Each participant received just the appearance verb texts from Fischer et al.'s 36 critical items, along with the attention check used in Experiment 1, in random order. For each item they had to indicate whether they regarded the protagonist mentioned in the text or the text's author as patient of the appearance verb. For instance, for text 6a (above) participants were asked 'Do you

[11] P-values for the significant items ranged from 0.00090 to machine error ($< 2.2e^{-16}$). For the remaining three items we found $t(105) = 1.5645$, $p = 0.12$ ('The girl looked Korean to Jack. Jack believed she was Japanese.'); $t(105) = 0.88119$, $p = 0.38$ ('Michael's socks looked dark blue to him. Michael thought they were black.'); $t(105) = 0.1241$, $p = 0.90$ ('Their efforts seemed idealistic to Sam. Sam thought they were self-serving.').

[12] 67.3% of participants gave a response above the neutral point across the visual items, 71.2% did so across the abstract items.

[13] Participants were 68.3% women, average age 42.3, ranging from 16 to 76.

interpret the first sentence in this text in terms of the rambler finding that the hill seemed quite steep or in terms of the author finding that the hill seemed quite steep?' Participants answered by selecting either 'The Rambler' or 'The Author'.

In line with h2, for each of the 36 items, a large majority of participants selected 'The Author', with the proportion ranging from a low of 75.0% to a high of 90.9%. Summing across the visual items, we found that a significant majority selected 'The Author' (82.8%) $\chi^2 = 340.1$, $p < 0.0001$. Results were almost identical for the abstract items with a significant majority selecting 'The Author' (83.0%) $\chi^2 = 342.73$, $p < 2.2e^{-16}$. Obviously, the proportions were not significantly different $\chi^2 = 1.1923e{-}29$, $p = 1$.

3.4 Experiment 3

To assess h3, we elicited subjectivity ratings for the claims under discussion.

We recruited 43 participants with the same approach and from the same population as in the previous experiments.[14] Each participant received just the first sentence from the 'is' texts for each of Fischer et al.'s 36 critical items, e.g.,

The hill was quite steep.

Items, along with the attention check used in Experiment 1, were presented in random order. For each item participants rated whether it expressed an objective fact or a subjective opinion on a scale from 1 ('completely objective') to 7 ('completely subjective').

The mean subjectivity rating varied notably across the items, ranging from a low of 2.23 ('the bird was a Hammerkop') to a high of 5.53 ('the young artist was talented'). The ratings varied between the visual and the abstract items. The mean of individual participants' average ratings for the visual items was significantly lower (M = 3.25, SD = 0.67) than for the abstract items (M = 4.40, SD = 0.67) $t(42) = -8.5395$, $p < 0.0001$. Further, the mean of the average ratings for the visual items was significantly below the neutral point $t(42) = -7.3164$, $p < 0.0001$, while the mean of the average ratings for the abstract items was significantly above the neutral point $t(42) = 3.9255$, $p = 0.00032$. In other words, participants tended to treat the visual sentences on average as being distinctly objective, the abstract sentences on average as being distinctly subjective, and the visual items as less subjective than the abstract items, as predicted by h3.

3.5 Discussion

The results of Exp.1 suggest that, in the critical items from Fischer et al.'s (2019) study, English-speakers generally find a patient-role assignment to the texts' protagonist to be contradictory, for both items with visual objects and for items with abstract objects. The significant but slight difference found in ratings for the two item types (abstract > visual) is consistent with h1. Crucially, the fact that items of both types were deemed distinctly contradictory continues to motivate the hypothesis h2 that participants will tend to assign the patient-role of appearance verbs in both types of items to

[14] Participants were 65.1% women (two non-binary), average age 38.5, ranging from 16 to 74.

the author. Exp.2 tested h2 more directly, using a transparent design that made the task of patient-role assignment explicit and suggested an alternative assignment (namely, to the author). Participants clearly preferred the alternative assignment. Nonetheless, while the results are suggestive, they only provide limited warrant for the conclusion that such assignment will also happen when task and alternatives remain implicit. The results of Exp.3 supported h3, indicating that participants tended to see the texts with visual objects as being less subjective than the texts with abstract objects. Together, Exp. 1–3 provide initial support for the alternative explanation and motivate examination of the hypothesis that Fischer et al.'s (2019) study is subject to a confound.

In its strongest, but plausibly most intuitive form, the hypothesis posits two correlations:

H_2 Participants largely base their plausibility assessments for the critical items on subjectivity-assessments for the claim under discussion: When the perceived subjectivity of this claim increases, appearance-sentences strike participants as increasingly appropriate, and the plausibility of appearance-texts (like 6a) increases; when the perceived subjectivity of the claim under discussion decreases (i.e., its perceived objectivity increases), 'is'-sentences strike participants as increasingly appropriate, and the plausibility of 'is'-texts (like 6b) increases.

As a result, participants in Fischer et al.'s study will tend to regard 'is'-texts as more plausible than otherwise identical appearance-texts when they feel the claims under discussion are relatively objective, and less plausible than appearance-counterparts when the claims under discussion feel relatively subjective. If so, this, rather than persistent inappropriate inferences from the appearance verbs, might explain the observed preferences.[15]

4 Main study

To more rigorously examine the key hypothesis H_1 thrown into doubt by the studies reported in the previous section, as well as to address H_2, our main study implemented the cancellation paradigm with a combination of eye tracking and plausibility ratings. By combining these two measures, we can separately test hypotheses about what automatic inferences are initially triggered (H_1, part 1) and whether they get suppressed or influence further cognition (H_1, part 2).

[15] This alternative explanation makes two assumptions: First, participants prefer the text with the higher individual plausibility rating. Second, if (as per H_2) plausibility assessments for individual sentences/texts are based on subjectivity ratings for the claim under discussion, preferences should be random, where subjectivity ratings are neutral (and therefore do not favour either 'is' or 'appear'). With H_2, higher (above neutral) subjectivity ratings could then explain preferences for appearance texts, and lower (below neutral) ratings preferences for 'is' texts. While H_2 posits two different correlations between subjectivity and objectivity ratings (positive for appearance texts, negative for 'is' texts), each of these correlations could explain preferences in the absence of the other. (By contrast, the alternative explanation cannot work, if both correlations go the same way, or no correlations obtain.) While intuitively plausible, H_2 thus is stronger than strictly required.

4.1 Predictions

In this study, participants read and rated items in which appearance sentences employing 'look', 'appear', or 'seem' were followed by a sequel that was either inconsistent or consistent with the hypothesised stereotypical inference from the appearance verb to a doxastic conclusion:

(1) The dress seemed blue. Hannah thought it was green. (*stereotype-inconsistent*)
(2) The dress seemed blue. Hannah thought it was navy. (*s-consistent*)

As in the study of Fischer and colleagues (2019), stereotype-inconsistent items with 'look', 'appear', and 'seem' were intended to invite phenomenal (re-)interpretation of the verb. Participants also read and rated otherwise identical items with the contrast verb 'is', which lacks stereotypical association with doxastic patient properties (where, for convenience, we retain the label stereotype- or 's-in/consistent' for counterparts of stereotype-in/consistent appearance items):

(3) The dress was blue. Hannah thought it was green. (s-inconsistent)
(4) The dress was blue. Hannah thought it was navy. (s-consistent)

When we read sentences, our eyes may pass over the same words several times.[16] Whereas *first-pass reading times*[17] are largely determined by word length, word frequency, and the word's predictability in context ('cloze probability') (Rayner 1998), difficulties in integrating information from different parts of the sentence may have us reread bits of the sentence (Rayner et al. 2004; Clifton et al. 2007). Specifically, such integration difficulties may have us reread the regions where the difficulty becomes manifest ('conflict region') and regions perceived as the source of the difficulty ('source region'). Where inferences triggered by previous words ('seemed blue') clash with subsequent text ('Hannah thought it was green'), this leads to higher rereading times for either the conflict region ('green') or the source region ('seemed blue'), or both. This increases the *total reading times* for these regions (defined as the sum of all fixations in a region) and the *second pass reading times* (defined as total minus first pass reading times). These two measures are known as 'late' reading times. The hypothesis that appearance verbs trigger doxastic inferences (H_1, part 1) predicts

[Prediction RT] Late reading times for conflict or source regions will be higher in s-inconsistent appearance-items (like 1 above) than in 'is'-counterparts (like 3 above).

When initially triggered stereotypical inferences clash with contextual information or with background beliefs, they can be suppressed within one second and before they influence further cognition (Fischer and Engelhardt 2017b). The hypothesis that the doxastic inferences of interest influence further cognition (H_1, part 2) therefore needs to be tested separately. It predicts that, in a subsequent non-speeded plausibility

[16] For an accessible introduction to reading times and eye tracking, see Raney et al. (2014), for current state of the art Clifton et al. (2016), for adaptation to study of reasoning Fischer and Engelhardt (2019b).

[17] These are defined as the sum of all fixations in a region of text, from first entering that region until leaving that region either in a forward or backward direction (Clifton et al. 2007). The terminology ('second pass reading times', etc., see below) is not uniform in the literature.

rating task, these inferences will reduce the plausibility of s-inconsistent appearance sentences, where they clash with the sequel, but will not affect the plausibility of s-consistent items. Hence:

[Prediction PL1] S-inconsistent appearance items (like 1 above) will be deemed less plausible than s-consistent appearance items (like 2 above).

Since s-inconsistent 'is'-items claim that protagonists are wrong about typically obvious matters (like the colour of a dress), we would expect participants to find them mildly implausible. However, the doxastic inferences posited by H_1 would render s-inconsistent appearance items outright contradictory, and reduce their plausibility even further. Hence:

[Prediction PL2] If s-consistent appearance and 'is' items (like 2 and 4) are deemed equally plausible, s-inconsistent appearance items (like 1) will be deemed less plausible than 'is'-counterparts (like 3).

Such plausibility differences would provide evidence of cognitively influential doxastic inferences from appearance verbs, in inappropriate contexts (namely, in s-inconsistent items which invite phenomenal interpretation of the word).

The competing Hypothesis H_2 suggests that plausibility judgments about s-inconsistent texts (used in the previous plausibility ranking study and figuring among the items in this new study) are driven by subjectivity ratings (rather than by contextually cancelled doxastic inferences). To assess this hypothesis, we elicit subjectivity-ratings for the claims under discussion in our items. These claims are expressed by the first sentences of 'is' versions (e.g., 'The dress was blue'). H_2 assumes that readers of s-inconsistent appearance items will assign the verb's patient role to the author. This assignment—and only this assignment—turns the appearance-sentence into the expression of an authorial opinion (*I think the dress was blue*) (Sect. 3.1). H_2 suggests that participants then find appearance sentences, interpreted as expressions of opinions, more appropriate, and appearance items more plausible, the more subjective they deem the claim under discussion, and will find corresponding 'is' sentences more appropriate, and 'is'-items more plausible, the more objective (less subjective) they deem the claims under discussion (Sect. 3.5). H_2 thus predicts:

[Prediction PL3] For appearance items, there will be a positive correlation between subjectivity ratings (for claims under discussion) and plausibility ratings (for items); for corresponding 'is' items, there will be a negative correlation between these ratings

In stereotype-inconsistent items, the patient-role assignment assumed by H_2 can be either due to reassignment in response to the perceived inconsistency or made from the start (Sect. 3.1). If that assignment is made from the start, H_2 should apply also to stereotype-consistent items. Since we have not ruled out this possibility, we will examine [PL3] first for all items and then for stereotype-consistent and -inconsistent items, separately.

4.2 Methods

4.2.1 Participants

Forty-eight first- and second-year undergraduate psychology students (9 males) from the University of East Anglia participated for course credit. All were native speakers of English with normal or corrected-to-normal vision.

4.2.2 Materials

Each participant read 48 critical items (six for each of eight conditions) and 48 fillers. All items were about visual objects. Half of the critical items involved basic visual properties (colour, shape, size, 8 items each). The other half involved less basic, but easily visually ascertainable properties like material (silver, wood), or age (young, old). S-inconsistent items used antonyms in first and second sentence. S-consistent items used synonyms, or the second sentence used a sub-ordinate category (blue—navy). Appendix 1 gives a list of critical items. As verbs were rotated across items, mean length and frequency of words in the source regions were the same across verb conditions (except for the unavoidable differences between 'is', 'look', 'appear', and 'seem'). Following the norming work (described below) we ensured that, in the conflict regions, neither the mean frequencies (consistent: 126, inconsistent: 182 occurrences in reference corpus Leech et al. 2001) nor the mean lengths (consistent: 5.54 characters, inconsistent: 5.25) of the adjectives differed significantly between the s-consistent and the s-inconsistent items (length: $t(46) = -0.58, p = 0.57$, frequency: $t(46) = 0.80$, $p = 0.43$).

To guard against floor effects and ensure intelligibility of items, a *norming study* with twenty-six participants from the same population rated the plausibility of 'is'-versions of candidate items (half s-consistent, half s-inconsistent), on a 5-point scale. Participants identified words they did not understand and did not rate the items containing them. We excluded all items where the s-inconsistent version attracted a mean rating < 2.5, and excluded or rephrased all items where at least two participants failed to understand a constituent word.

4.2.3 Apparatus

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

4.2.4 Design and procedure

We manipulated the verb in the first sentence ('is', 'look', 'appear', 'seem') and the consistency of the sequel with hypothesised doxastic inferences (s-consistent vs. s-inconsistent), in a 4×2 design. All variables were manipulated within subject. We

measured first pass, second pass, and total reading times for source regions, conflict regions, and their constituent words.

After a 9-point calibration and validation procedure, participants completed two practice trials and 96 experimental trials. These included 48 critical trials. Each participant saw an equal number of items in each condition, as verbs were rotated across items using a Latin Square Design. Before each trial, participants fixated a drift-correction dot on the left edge of the monitor, centred vertically. The sentence appeared after an interval of 500 ms. The initial letter of each sentence was displayed in the same position as the drift correction dot. The entire sentence appeared on a single line on the screen. The participant read the sentence silently and then pressed the spacebar on the keyboard. A plausibility-rating prompt appeared, and participants rated sentences' plausibility on a scale from 1 to 5, by pressing the corresponding key on the keyboard. Endpoints were explained as 'very implausible' (1) and 'very plausible' (5), and the midpoint (3) as 'neither plausible nor implausible; the decision feels arbitrary'.

4.3 Results

To preview findings, results largely bore out predictions derived from H_1, but not predictions from H_2.

We analysed *plausibility ratings* for all items with a 2 × 4 (context × verb) repeated-measures ANOVA. This revealed large main effects of consistency $F(1,46) = 387.21$, $p < 0.001$, $\eta^2 = 0.89$ and verb $F(1,46) = 7.53$, $p < 0.01$, $\eta^2 = 0.14$, and a marginal 2-way interaction $F(1,46) = 3.29$, $p = 0.076$, $\eta^2 = 0.07$ (see Fig. 1). Participants rated s-consistent items distinctly plausible, or significantly above neutral mid-point, in all verb conditions (p's < 0.001 for all mean ratings), and deemed s-consistent items with different verbs equally plausible $F(3,138) = 0.59$, $p = 0.62$, $\eta^2 = 0.01$. By contrast, s-inconsistent items with all verbs were deemed distinctly implausible, or significantly below mid-point (all p's < 0.001), and there were significant differences between verb conditions $F(3,138) = 3.54$, $p < 0.05$, $\eta^2 = 0.07$. As per prediction [PL1], s-inconsistent items with an appearance verb were deemed less plausible than s-consistent counterparts ('look': $t(46) = 15.07$, $p < 0.001$; 'appear': $t(46) = 15.87$, $p < 0.001$; 'seem': $t(46) = 16.11$, $p < 0.001$). Prediction [PL2] predicted that if s-consistent items with appearance verbs and 'is' are deemed equally plausible, the consistency manipulation will render appearance items less plausible than 'is' items. Participants indeed rated s-consistent items with all verbs equally plausible (above). As predicted by [PL2], s-inconsistent items with 'appear' and 'seem' were deemed less plausible than s-inconsistent items with the contrast verb 'is' (appear vs. is: $t(46) = -2.09$, $p = 0.04$; seem vs. is: $t(46) = 2.65$, $p = 0.01$). The mean plausibility rating for s-inconsistent items with 'look' (2.40) was numerically lower than the mean rating for similar 'is'-items (2.43), but, against predictions, this difference was not significant $t(46) = 0.36$, $p = 0.72$. Further comparisons between s-inconsistent items revealed that the difference between 'appear' and 'seem' was not significant $t(46) = 0.74$, $p = 0.46$, the difference between 'look' versus 'seem' was significant $t(46) = 2.48$, $p = 0.02$, and that between 'look' versus 'appear' trended towards significance $t(46) = -1.41$, $p = 0.09$.

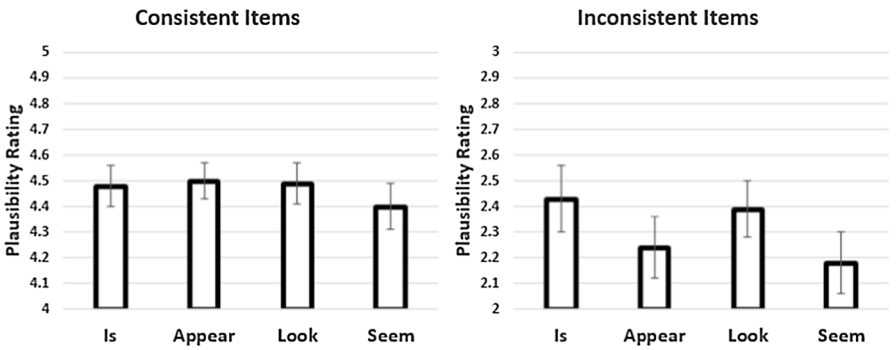

Fig. 1 Mean plausibility ratings. Error bars show the standard error of the mean

Our prediction [RT] about *reading times* only concerns later reading times for s-inconsistent items. The observed plausibility ratings suggest that later reading times will vary between verbs only for such items. Our analysis of reading times therefore focused on s-inconsistent items. We predicted that late (total and second pass) reading times for conflict or source regions will be higher for s-inconsistent items with verbs 'look', 'appear', and 'seem' than for corresponding items with 'is'. However, most relevant trials involved a striking pattern of eye movements: When reading s-inconsistent items, participants regressed from the end of the final sentence to the source region (e.g., to 'seemed blue' in 'The dress seemed blue. Hannah thought it was green'), reread the source region, and then progressed to the plausibility-rating screen without rereading the conflict region (e.g., 'green'). To take these findings into account, we report total reading times for the conflict region. Since second pass (= total minus first pass) reading times are the most precise measure of integration difficulties, we report these for the crucial source region.

We analysed reading times for s-inconsistent items with a one-way repeated-measures ANOVA with verb type having four levels (manipulated within item). This revealed that total reading times for the conflict region (e.g., 'green') were not significantly different for items with different verbs $F(3,138) = 1.39$, $p = 0.25$, $\eta^2 = 0.03$ (see Fig. 2). By contrast, second pass reading times for the source region, obtained by summing across the first verb and first object (e.g., 'seemed blue'), showed a large effect of verb $F(3,138) = 4.65$, $p < 0.01$, $\eta^2 = 0.24$. Paired comparisons revealed second-pass reading times were appreciably higher in 'appear'-items than 'is'-items $t(46) = 3.41$, $p = 0.001$ and in 'seem'-items than in 'is'-items $t(46) = -2.85$, $p = 0.007$, consistent with prediction [RT]. By contrast, the difference between 'look'-items and 'is'-items was not significant $t(46) = -1.34$, $p = 0.19$. Differences between items with different appearance verbs also remained shy of significance: Differences were marginally significant between 'look'- and 'appear'-items $t(46) = 1.92$, $p = 0.06$, but not between 'look'- and 'seem'-items $t(46) = -1.55$, $p = 0.13$. For further reading times, with discussion, see Appendix 2.

As we will presently discuss, these findings provide evidence that the intended phenomenal uses of 'appear' and 'seem' in our items triggered doxastic inferences

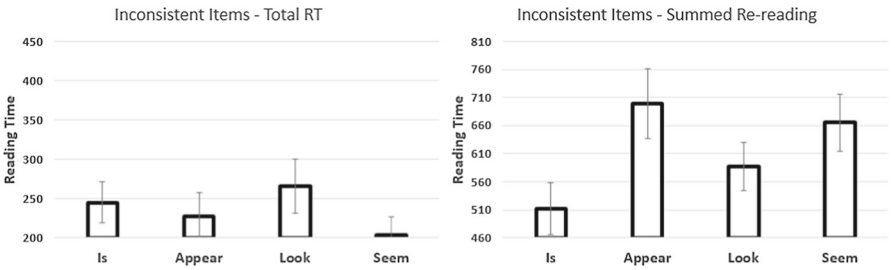

Fig. 2 Left panel shows total reading time on the *conflict region*. Right panel shows the summed second-pass reading time on verb and adjective jointly making up the *source region*. Error bars show the standard error of the mean

that influence further cognition, and support hypothesis H_1 for 'appear' and 'seem', but not for 'look'.

4.4 Follow-up study

To assess whether plausibility ratings were largely based on subjectivity judgments, in the way suggested by H_2, we recruited 107 participants with the same approach and from the same population as in Experiments 1–3.[18]

Participants rated the claims under discussion for all critical items used in the main study, as expressed by the first sentence of items' 'is' versions (e.g., 'The dress was blue'). Participants rated them on a scale from 1 ('completely objective') to 7 ('completely subjective'). We then calculated the mean subjectivity ratings and reanalysed the data from the main study to assess H_2's prediction [PL3] that mean plausibility ratings for items would correlate with mean subjectivity ratings for claims under discussion, positively for items with appearance verbs and negatively for items with 'is'. Accordingly, the following analyses were conducted on items, rather than participants.

Mean subjectivity ratings for claims under discussion varied between 1.52 ('The word's spelling was correct') and 5.31 ('The building was quite grand'), and neatly divided into halves close to mid-point (4), with half the claims receiving mean ratings of 4.2 or above. A two-way ANCOVA, which included context (consistent and inconsistent) and verb type (with four levels) and subjectivity ratings as a covariate, showed significant main effects of context $F(1,22) = 16.87$, $p < 0.001$, $\eta^2 = 0.43$ and subjectivity $F(1,22) = 5.78$, $p = 0.025$, $\eta^2 = 0.21$ as well as a significant interaction between context and subjectivity $F(1,22) = 5.13$, $p = 0.03$, $\eta^2 = 0.19$. Crucially, however, results revealed no significant effect of verb $F(1,22) = 1.29$, $p = 0.27$, $\eta^2 = 0.001$ and no interaction between verb and subjectivity $F(1,22) = 0.15$, $p = 0.74$, $\eta^2 = 0.001$, and there were also no significant correlations between subjectivity and plausibility ratings (all p's > 0.42). This is inconsistent with prediction [PL3] derived from hypothesis H_2.

[18] Participants were 83.2% women (one non-binary), average age 40.2, ranging from 16 to 77.

However, H_2 was initially advanced as a hypothesis about s-inconsistent items (Sect. 3.4), and was then tentatively extended to s-consistent items (Sect. 4.1). Next, we therefore considered stereotype-consistent and—inconsistent items separately. The s-consistent items showed no significant main effects or interaction (all p's > 0.30). The correlations between subjectivity and plausibility ratings were not significant, either (is: $r = 0.07$, $p = 0.75$; look: $r = 0.16$, $p = 0.45$, appear: $r = 0.05$, $p = 0.83$; seem: $r = -0.08$, $p = 0.70$). That is: Differences in subjectivity did not correspond to any changes in plausibility. The more subjective half of the s-consistent items were deemed roughly as plausible as the more objective half of the s-inconsistent items, across all verb conditions (subjective half mean plausibility = 4.51 and objective half mean plausibility = 4.41).

The crucial s-inconsistent items showed no main effect of verb $F(1,22) = 2.78$, $p = 0.11$, $\eta^2 = 0.11$, but a significant and large main effect of subjectivity $F(1,22) = 14.40$, $p = 0.001$, $\eta^2 = 0.40$. Crucially, however, there was no interaction between verb and subjectivity $F(1,22) = 1.10$, $p = 0.31$, $\eta^2 = 0.048$. Against prediction [PL3], we found negative correlations between subjectivity and plausibility ratings not only for items with 'is' ($r = -0.50$, $p = 0.014$) but also for items with 'look' ($r = -0.56$, $p = 0.005$), 'appear' ($r = -0.47$, $p = 0.019$), and 'seem' ($r = -0.44$, $p = 0.030$). That is: The more subjective the relevant claims were deemed, the less plausible s-inconsistent items were judged, regardless of the verb used (i.e., there was no effect of verb). The more subjective half of the s-inconsistent items were deemed less plausible than the more objective half of the s-inconsistent items, across all verb conditions (subjective half mean plausibility = 2.07 and objective half mean plausibility = 2.53). Strikingly, plausibility ratings were numerically higher for 'is' items (2.17) than for appearance items (2.04), for the more subjective half of our s-inconsistent items (where H_2 suggested that appearance verbs would be deemed more appropriate and appearance items more plausible than items with the supposedly more objective 'is').

4.5 Discussion

The follow-up study's findings speak against H_2, and against subjectivity being a confound for the main study. According to H_2, participants base plausibility ratings for items on how appropriate the first sentence's verb ('is' vs. 'appear', etc.) is in view of the subjectivity of the claim under discussion. Differences in plausibility are then attributed to differences in subjectivity, which are held to affect the plausibility of 'is' and appearance sentences in different ways. However, when considering the whole sample, we found no significant correlations between subjectivity ratings (for claims under discussion) and plausibility ratings (for items), for any verb condition—despite a main effect of subjectivity. Moreover, this effect disappeared when we considered only s-consistent items (which employ the same first sentences as s-inconsistent items). This suggests that participants' plausibility ratings were not influenced mainly by the first sentence and the fit between its verb and the subjectivity of the claim under discussion. Subjectivity ratings did affect the plausibility of items in the s-inconsistent condition. However, it affected the plausibility of all s-inconsistent items in the same way, namely decreased it across all verb conditions—rather than

increasing the plausibility of appearance items and decreasing that of 'is' items (as per H_2). Differences in subjectivity therefore cannot explain the differences in plausibility between s-inconsistent items with different verbs.

The follow-up findings help us interpret the main study's findings concerning H_1. They help address the question of how participants interpret the crucial s-inconsistent items in the main study ('The dress seemed blue. Hannah thought it was green'): whether they assign the patient role of the appearance verb to the protagonist (Hannah) or to the author, either from the start or as a result of reassigning the patient role from protagonist to author in response to the otherwise severe clash with the sequel (see Sect. 3.1). Assignment of the patient role to the author would turn the appearance sentence into the expression of an authorial opinion or hedged judgment about a matter of fact. This would lead participants to interpret s-inconsistent items as expressing a clash of subjective opinions between the author and the protagonist. The claim that there is such a clash should seem more plausible when the matter at hand is deemed more subjective. In line with [PL3], patient-role assignment to the author thus predicts a positive correlation between subjectivity and plausibility ratings for appearance-items. However, we observe a negative correlation for the crucial s-inconsistent items. We infer that, as long as no alternative assignment is explicitly suggested to them (as in Exp.2), participants assign the patient-role of the appearance verb to the protagonist in the text (*The dress seemed blue to Hannah*).

As long as this assignment is maintained, the most obvious interpretation of s-inconsistent items that avoids a contradiction is the phenomenal interpretation intended by the experimenters. On this reading, the items' first sentence attributes merely an experiential attitude to the protagonist (The dress visually appeared blue, to Hannah), but no doxastic attitude (Hannah may well still believe that it has another colour than it looks to her, here and now, under these lighting conditions, etc.). This avoids contradiction with the sequel but makes for a mildly implausible scenario. Since the items do not make any deviations from stereotypical viewing situations explicit, participants will infer with the I-heuristic (Levinson 2000; see Sec. 2.1 above) the absence of factors (like odd lighting) that could lead a protagonist to distrust appearances, and will think it mildly implausible that the protagonist should think the object has one visual property (as asserted by the sequel), when it looks another to them. To win through to this non-contradictory, phenomenal, interpretation of these implausible items, readers need to completely suppress the doxastic component features of the situation schemas associated with appearance verbs.

Higher second-pass reading times for source regions in s-inconsistent appearance sentences ('seemed green') than in corresponding 'is' sentences ('was green') could be due either to such suppression effort or to cognitive effort expended on patient reassignment (from protagonist to author, see above). Having excluded patient reassignment, we interpret observed elevated second-pass reading times for source regions in sentences with 'appear' and 'seem' as evidence of the effort to suppress contextually inappropriate schema components that is involved in phenomenal interpretation of the appearance verb.[19] Lower plausibility ratings for s-inconsistent items with 'seem' and

[19] Discussion of the technical psycholinguistic issue of why clashes in s-inconsistent items led to rereading primarily of source regions but not of conflict regions is beyond the remit of this paper.

'appear' than 'is' suggest that this effort meets with only partial success, and the doxastic conclusions inferred continue to influence ratings. We thus take these findings to support hypothesis H_1 for two of the three appearance verbs examined: At any rate phenomenal uses of 'seem' and 'appear' trigger doxastic inferences which are at most partially suppressed and continue to influence further judgment.

In contrast with 'appear' and 'seem', 'look'-items pattern with 'is'-items, in the s-inconsistent condition: Neither the second-pass reading times for the source regions nor the plausibility ratings are significantly different, and both are deemed distinctly implausible (if more plausible than 'appear' and 'seem'). This suggests that participants construct for both 'look' and 'is'-sentences situation models where the protagonist *looks at* the object; they then find it roughly equally implausible that the protagonist should judge the object viewed to have one property, when it actually possesses ('is') or *looks* another, in stereotypical viewing conditions (see above). Nearest neighbour analyses of 'look at' (Fischer et al. 2019; Fn.22) and 'look F' (Fischer et al. 2015, Appendix) suggest these verbs are associated with epistemic and doxastic features, respectively, but more weakly than 'seem' and 'appear' are. In both cases, these stereotypical features need to be suppressed, to arrive at a consistent—e.g., phenomenal—interpretation of s-inconsistent items. Given the weaker association of the relevant features, suppression requires less effort than with 'seem' and 'appear', as evidenced by lower second-pass reading times for the source region, and this lesser suppression effort meets with greater suppression success, as evidenced by higher (if still low) plausibility ratings in the main study. The fact that, in Exp.1, 'look' was the only appearance verb which led to lower contradictoriness ratings for s-inconsistent items with visual objects than abstract objects provides further evidence that participants find it easier for 'look' than 'appear' and 'seem' to win through to a largely phenomenal interpretation which attributes experiential features (available for items with visual objects, but not with abstract objects) but is largely devoid of doxastic implications.

Yet further evidence is provided by a follow-up study we undertook in response to a reviewer query (and report in Appendix 3). Participants judged the acceptability of items that make, respectively, doxastic and non-doxastic uses of appearance verbs to describe familiar cases of non-veridical perception (where nobody is taken in). For all three appearance verbs, they deemed non-doxastic uses acceptable (e.g., 'Seen from the beach, the huge ships anchored out at sea look small', though nobody believes they are small). This suggests that all three verbs are used—also—in a non-doxastic 'phenomenal' sense, in ordinary discourse. But non-doxastic uses were deemed more acceptable for 'look' than the other verbs. This suggests that for 'look' this non-doxastic phenomenal sense is more salient, and the doxastic sense less salient, than for the other verbs, so that contextually inappropriate doxastic inferences are easier to suppress, and exert less influence on further cognition (as per the Salience Bias Hypothesis).

To sum up, present findings largely confirm H_1 and suggest that phenomenal uses of 'seem' and 'appear' trigger contextually inappropriate doxastic inferences that influence further cognition, whereas any doxastic inferences readers may make from 'look' can be swiftly suppressed. The main study also helps us assess the hypothesis H_2 that explains observed plausibility judgments with reference to subjectivity judgments

rather than to inappropriate doxastic inferences. Present findings speak against this potential confound affecting the present main study. It also excludes this confound for a previous plausibility ranking experiment (Fischer et al. 2019) (see Appendix 4).

However, present findings partially diverge from findings of this and another previous study employing the forced-choice plausibility-ranking task: Fischer et al. (2019) observed significant preferences of 'is'-sentences over appearance sentences, in s-inconsistent items with all three appearance verbs. In the present study, mean plausibility ratings were numerically minimally lower for s-inconsistent 'look'- than for 'is'-sentences. The forced-choice plausibility-ranking paradigm may translate such insignificant plausibility differences into significant differences in preference. However, an earlier study (Fischer and Engelhardt 2016) using the same paradigm found random preferences for 'look'- over 'appear'-sentences, whose mean plausibility ratings in the present study display a larger numerical difference than 'look' and 'is'. The experimental findings specifically for 'look' thus present a mixed picture.

5 Main findings and philosophical application

This paper's main study provided evidence that competent language users make contextually inappropriate stereotypical inferences from phenomenal uses of appearance verbs 'appear' and 'seem' to attributions of doxastic attitudes, which influence further judgment even when explicitly cancelled by the context. In conjunction with prior distributional semantic analysis (Sect. 2.3), these findings suggest that, in their intransitive use, 'appear' and 'seem' are primarily and irrepressibly doxastic terms; they are primarily used to attribute doxastic attitudes to patients and cannot completely shed their doxastic implications. By contrast, 'look' is used less often for this purpose and can be more easily interpreted as simply commenting on the looks of a thing, rather than the doxastic attitudes of its beholder (*cf.* Austin 1962, pp. 36–37).

5.1 Re-assessing the target argument

These findings help us assess the influential 'argument from illusion', whose initial premise typically uses 'appear' or 'seem' (rather than 'look') to describe cases of non-veridical perception (Sect. 1.1). Present findings allow us to support the key elements of our initial a priori reconstruction of the argument (Sect. 1.1) with an experimentally supported empirical explanation (*cf.* Sect. 1.2).

The explanation is this: In line with the Salience Bias Hypothesis (Sect. 2.2), inappropriate doxastic inferences lead from the phenomenal use of appearance verbs in the initial premise (1) 'When viewed sideways, the round coin appears elliptical' to the implicit conclusion (C) 'The viewer thinks that the object viewed is elliptical.' Integration of (C) with the contextual information that the coin is round leads to the conclusion (C*) 'The viewer has a wrong belief about the coin, and does not know that it is round, or that there is a round coin.' These conclusions feed into the situation model. The categorization question, whether viewer and object fall under the category *x is aware of y*, is then assessed on the basis of conformity with the 'aware-

ness' stereotype, with a version of the representativeness heuristic (Kahneman and Frederick 2002). The most highly weighted component features of the 'awareness' stereotype are epistemic (Fischer et al. 2019), so (C*) suggests conformity is low, and the representativeness heuristic delivers the judgment (2) that (more likely than not) the viewer is not aware of the round coin. Together, the Salience Bias Hypothesis and a well-researched judgment heuristic can thus explain the move from the initial premise to the first explicit conclusion of the argument from illusion. In the face of plausible principles of charity (Thagard and Nisbett 1983) (Sect. 1.2), this empirically supported explanation justifies the diagnostic hypothesis that this influential argument is vitiated by a contextually inappropriate stereotypical inference from 'appears' or 'seems' that prevents the argument from getting off the ground (Sect. 1.1).[20,21]

So far, we have considered only the early twentieth century version of the argument from illusion, addressed by Austin (1962). *Prima facie*, the contemporary version (rendered canonical by Robinson 1994, and Smith 2002) seems to dispense with the move we explained. It invokes instead the Phenomenal Principle (ii below) and Leibniz' Law (iv below):

(i) When subjects view a round coin sideways, the coin appears elliptical to them.
(ii) Whenever something appears a sensible property F to observers, they are (directly) aware of something that actually has this property. Hence:
(iii) When subjects view a round coin sideways, they are (directly) aware of something that actually is elliptical (an elliptical speck).
(iv) If b has a property a lacks, $a \neq b$.
(v) When subjects view a coin sideways, they are (directly) aware of something other than the round coin (an elliptical speck or 'sense-datum').

The Phenomenal Principle, however, 'is nowhere near the truth. No one would agree that if someone looks old then there must be something old' (Snowdon 2015, p. 128). Fischer et al. (2019) suggest that the principle only seems compelling to proponents

[20] In the rare cases where 'look' is used in the initial premise, any inappropriate doxastic inferences triggered by the verb will not influence the further argument. In these cases, we submit, the argument is accepted due to belief bias (Thompson and Evans 2012) and prior belief in the conclusion, accepted due to standard versions of the argument or parallel arguments (e.g., arguments from hallucination).

[21] In line with the 'reflection defence' against challenges from evidential experimental philosophy (see Machery 2017, for a review), a reviewer questioned whether this explanation applies to philosophers who spent more time and attention considering the argument from illusion than our lay participants spent on processing our experimental items. Empirical studies have investigated how (a) response time, (b) environmental factors conducive to slow and careful reflection (increased response delay, financial incentives, demand for justification, and analytic priming), and (c) a disposition to engage in reflection influence verdicts about verbally described cases in philosophical thought experiments (Colaço et al. 2018; Weinberg et al. 2012). All these factors facilitate correction of less reflective judgments, where these are [i] explicit and [ii] can be corrected by using normative rules like those of arithmetic, logic, or probability theory. However, none of the factors investigated were found to influence case verdicts about verbally described cases in philosophical thought experiments. This matters here: The initial premise of arguments from illusion is a brief verbal case description. The argument's base step articulates a philosophical thought experiment. The crucial conclusions (C) and (C*) remain implicit and it is at best unclear whether normative (rather than heuristic) rules are available for their correction. Increased time and attention are therefore unlikely to prevent (C) and (C*) from influencing further judgment, and thus unlikely to change spontaneous case verdicts (like 2). Our account is thus likely to apply to philosophers too. We suspect that in case-based reasoning reflection is directed less at modifying than at justifying unreflective case verdicts (*cf.* Schwitzgebel and Ellis 2017).

of the argument because they misinterpret supposedly noncommittal talk of 'elliptical specks', etc.: In well-established metaphorical usage, these phrases are ordinarily used to pick out physical objects by the way they look to the speaker, there and then, when one cannot tell what they are ('What is that small red speck in the valley? Could that be a fire truck?') or wishes to avoid stereotypical implications of knowledge ('She saw the grey specks grow larger. Had she recognised them as enemy planes, she would have run for cover'). Proponents of the argument misinterpret these phrases as attributing properties to something because they already presuppose the conclusion (2 above) that the viewer is not aware of the physical objects she looks at. This now implicit conclusion renders the above metaphorical interpretation impossible, and leads to default literal interpretation, which in turn requires positing an alternative object of awareness as literal bearer of the property in question.[22] Also in its current version, the argument therefore relies on the spontaneous inference from the initial case description (1) to the negative conclusion (2) explicitly acknowledged by the earlier version of the argument—and is, on the account we propose, reliant upon the inappropriate stereotypical inferences from appearance verbs we documented and explained as a result of salience bias.

This paper's experimental demonstration of these inferences thus completes the first successful experimental implementation of critical ordinary language philosophy in the wake of Austin (1962): the exposure of a contextually inappropriate stereotypical inference at the root of an influential philosophical argument, namely, the argument from illusion that was already Austin's main target.

5.2 Directions for future research

The argument from illusion is arguably not the only influential philosophical argument reliant on inappropriate stereotypical inferences that result from salience bias. Philosophers often take words that already have a dominant sense in ordinary discourse and use them in a related but rare—or even entirely new—sense, to meet specific philosophical research needs or to talk about unusual cases. Wherever this happens, reliance on a common polysemy processing strategy (Retention/Suppression strategy, Sect. 2.2) will lead to salience bias and inappropriate inferences that may vitiate philosophical argument—confirming Austin's (1962, p. 63) hunch that 'tampering with words … is always *liable* to have unforeseen repercussions.'

Thus, it has been suggested that arguments from hallucination (e.g., Ayer 1956; Smith 2002) rely on factive inferences that are licensed by the dominant perceptual sense of perception verbs, but not the phenomenal sense intended in the arguments (Fischer and Engelhardt 2019a) and that Chalmers' (1996) zombie argument relies on inferences from his technical use of 'zombie' that are licensed only by the noun's dominant ('Hollywood') sense and are defeated by contextual information essential to

[22] The plausibility of this literal interpretation may turn on implicit adherence to the Cartesian Theatre conception of the mind as a complementary space of perception, which can house these alternative objects of awareness (Fischer et al. 2015, p. 286). This implicit theory may thus support accommodation of the inappropriate stereotypical inference that renders plausible 'phenomenal' judgments like (iii) above. The Phenomenal Principle is then formulated to rationalize these judgments.

the argument (Fischer and Sytsma 2020).[23] At the same time, evidence that the ordinary use of causal attributions is morally laden (e.g., Livengood et al. 2017; Livengood and Sytsma 2020), whereas metaphysical argument often intends a purely descriptive use, motivates the question whether common philosophical thought experiments that use morally valenced cases (assassinations, etc.) can support metaphysical arguments about 'causation' in the intended descriptive sense.

As developed by Austin (1962), critical ordinary language philosophy seeks to expose 'seductive (mainly verbal) fallacies' that may act as 'concealed motives' for adopting persuasive but unwarranted philosophical conclusions that clash with background knowledge—and thus generate bogus problems. Following Austin's lead, this paper examined contextually inappropriate stereotypical inferences from words that thinkers cannot help going along with, even when they explicitly reject them. However, not only automatic comprehension inferences we renounce, but also implicit theories we explicitly reject, may lead to clashes between conclusions of implicit cognition and explicit knowledge. Thus, implicit extramissionist theories of vision—according to which something 'leaves the eye' when we see—have been experimentally found to influence judgments about perceptual objects, even when thinkers explicitly reject those theories (Guterstam et al. 2019; *cf.* Shtulman and Valcarcel 2012). The two factors may work together in philosophical argument. For example, an implicit theory of the mind as an inner space of perception may facilitate accommodation of inappropriate stereotypical inferences, in the argument from illusion (Fn.22). This suggests exciting new perspectives for an experimental implementation of ordinary language philosophy's critical project that seeks to 'dissolve' philosophical problems by exposing, more generally, divergences between implicit cognition and explicit knowledge, arising from a variety of sources involving different kinds of conceptual structures (stereotypes and implicit theories).

With relevance beyond ordinary language philosophy, this paper explored methods that allow experimental philosophers to extend their investigation from intuitive judgments to automatic inferences. Experimental philosophers tend to content themselves with the use of single output measures. Psycholinguists tend to study automatic comprehension inferences with process measures. The present paper demonstrated that, as it stands, neither approach is fully satisfactory for the new purpose of studying natural language reasoning, and to exclude potential confounds: To examine with the psycholinguistic cancellation paradigm whether initially triggered comprehension inferences go on to influence further judgment and reasoning, we found it helpful to complement a process measure with plausibility ratings; to exclude a potential confound, we examined their correlation with another output measure. This model may be helpful for further study of how automatic comprehension inferences shape automatic inferences in natural language reasoning—including philosophical arguments and thought experiments.

[23] Using a different theoretical framework, Nichols and Pinillos (2018) propose that sceptical arguments are facilitated by inferences supported by an infallibilist concept of knowledge, that children acquire this concept through exposure to child-directed speech in which uses of 'know' consistent with infallibilism are dominant, and that this concept will influence reasoning even where 'know' is used in a fallibilist sense.

Acknowledgements This paper grew from a *Brains Blog* symposium on experimental ordinary language philosophy. We thank the symposium organizer Keith Allen and the *Brain Blog*'s Managing Editor, John Schwenkler, for bringing about this event. For previous comments on earlier drafts and closely related material, we are indebted to Keith Allen, Joachim Horvath, Pendaran Roberts, an interdisciplinary audience at the workshop 'Reasoning, Argumentation and Logic in Natural Language: Experiments and Models', Bochum, April 2019, and to two anonymous reviewers. For assistance with data collection for the main study, we thank Georgia Brown. Experiments 1–3 were supported by a University Research Fund grant from the Victoria University of Wellington (#220861).

Compliance with ethical standards

Ethics statement The research conformed to the ethical standards for conducting research as outlined by the British Psychological Society. The use of human research participants was approved by the relevant Research Ethics Committees of the University of East Anglia and of the Victoria University of Wellington.

Appendix 1: List of critical items

Verbs ("is", "look"; "appear", "seem") were rotated across lists in a Latin Square Design. Asterisks indicate regions of interest.

1 *The dress*was*blue.*Hannah*thought it was*green.*
2 *The dress*was*blue.*Anna*thought it was*navy.*
3 *The cake's icing*seemed*rose-coloured.*Tim*believed it was*white.*
4 *The cake's icing*seemed*rose-coloured.*Tom*believed it was*pink.*
5 *The socks*looked*dark blue.*Michael*thought they were*black.*
6 *The socks*looked*dark blue.*Andrew*thought they were*indigo.*
7 *The colours*appeared*bright.*Greg*believed they were*dull.*
8 *The colours*appeared*bright.*Chris*believed they were*lively.*
9 *The hill*was*steep.*Peter*thought it was*gentle.*
10 *The hill*was*steep.*Paul*thought it was*sheer.*
11 *The courtyard*seemed*rectangular.*Phil*believed it was*square.*
12 *The courtyard*seemed*rectangular.*Bill*believed it was*oblong.*
13 *The medal*looked*elliptical.*Joe believed it was*round.*
14 *The medal*looked*elliptical.*Jim*believed it was*oval.*
15 *The chef*appeared*heavy.*Sam*believed he was*slim.*
16 *The chef*appeared*heavy.*Sue*believed he was*fat.*
17 *The cat darting across the street*was*small.*Ben*believed it was*large.*
18 *The cat darting across the street*was*small.*Matt*believed it was*little.*
19 *The building*seemed*quite grand.*Alex*thought it was*humble.*
20 *The building*seemed*quite grand.*Will*thought it was*stately.*

21 *The dog partially hidden by the fence*looked*massive.*Daniel thought it was*small.*
22 *The dog partially hidden by the fence*looked*massive.*William*thought it was*large.*
23 *The bathroom*appeared*reasonably large.*Julian*believed it was*small.*
24 *The bathroom*appeared*reasonably large.*Gabriel*believed he was*big.*
25 *The shop*was*shut.*Sophie*believed it was*open.*
26 *The shop*was*shut.*Ellie*believed it was*closed.*
27 *The fruit*seemed*overripe.*John*believed it was*fresh.*
28 *The fruit*seemed*overripe.*Ron*believed it was*rotting.*
29 *The jacket*looked*much used.*Edward*thought it was*new.*
30 *The jacket*looked*much used.*Gareth*thought it was*old.*
31 *The deer lying at the roadside*appeared*conscious.*Amy*believed it was*dead.*
32 *The deer lying at the roadside*appeared*conscious.*Kim*believed it was*alive.*
33 *The dog in the yard*was*gentle.*Grace*thought it was*dangerous.*
34 *The dog in the yard*was*gentle.*Sarah*thought it was*harmless.*
35 *The office*seemed*disorderly.*Dan*thought it was*tidy.*
36 *The office*seemed*disorderly.*Dick*thought it was*messy.*
37 *The word's spelling*looked*correct.*Olivia*believed it was*wrong.*
38 *The word's spelling*looked*correct.*Emily*believed it was*right.*
39 *The man*appeared*advanced in years.*Mary*thought he was*young.*
40 *The man*appeared*advanced in years.*Emma*thought he was*old.*
41 *The group's actions*were*organised.*James*thought they were*haphazard.*
42 *The group's actions*were*organised.*Joshua*thought they were*coordinated.*
43 *The statue*seemed*bronze.*Michael*thought it was*gold.*
44 *The statue*seemed*bronze.*Andrew*thought it was*brass.*
45 *The cutlery*looked*silver.*Jack*believed it was*steel.*
46 *The cutlery*looked*silver.*Joe*believed it was*sterling.*
47 *The flooring*appeared*wooden.*Sarah*thought it was*laminate.*
48 *The flooring*appeared*wooden.*Eve*thought it was*mahogany.*

Despite our norming efforts, two s-consistent items (numbers 10 and 14) and two s-inconsistent items (numbers 5 and 45) attracted mean plausibility ratings that were outliers (SD's > 3) and were excluded from further analyses (i.e. plausibility and eye movements).

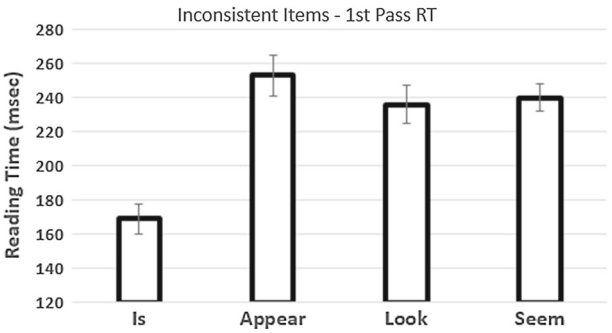

Fig. 3 Mean first pass reading times for the verb region in inconsistent items. Error bars show the standard error of the mean

Appendix 2: Further reading times

First pass reading times

Verb region

Results showed a significant effect of verb $F(3,138) = 19.18$, $p < 0.001$, $\eta^2 = 0.29$ (see Fig. 3). Paired comparisons revealed reading times were appreciably lower in 'is'-items than appearance verb items (appear-is: $t(46) = 5.88$, $p < 0.001$, look-is: $t(46) = -5.69$, $p < 0.001$, seem-is: $t(46) = -6.35$, $p < 0.001$). In contrast, the appearance items were not significantly different from one another (all p's > 0.18). The difference between 'is'-items and appearance items is easily explained by the shorter length and greater frequency of 'is' (Rayner 1998).

First object

Results showed no main effect of verb $F(3,138) = 0.92$, $p = 0.43$, $\eta^2 = 0.02$ (see Fig. 4). This means first pass reading times for the first object (e.g., 'blue' in item 1) were not significantly different for items with the contrast verb 'is' than for the appearance verbs.

Conflict region

Again, there was no main effect of verb $F(3,138) = 0.75$, $p = 0.52$, $\eta^2 = 0.02$ (see Fig. 5). This means first pass reading times for the conflict region (e.g., 'green' in item 1) were not significantly different for items with the contrast verb 'is' than for the appearance verbs.

Total reading time: source region

Total reading times on the source region (e.g., 'was blue', in item 1) showed a significant main effect of verb $F(3,138) = 9.22$, $p < 0.001$, $\eta^2 = 0.17$ (see Fig. 6).

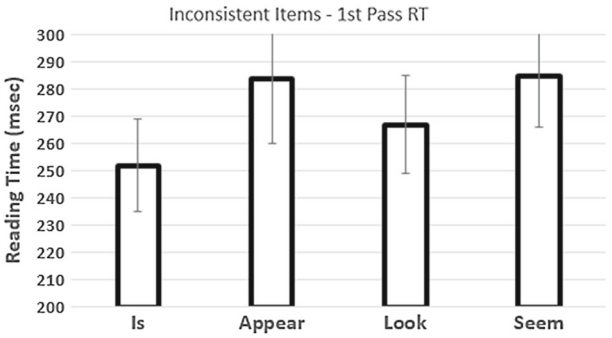

Fig. 4 Mean first pass reading times for the first object in inconsistent items. Error bars show the standard error of the mean

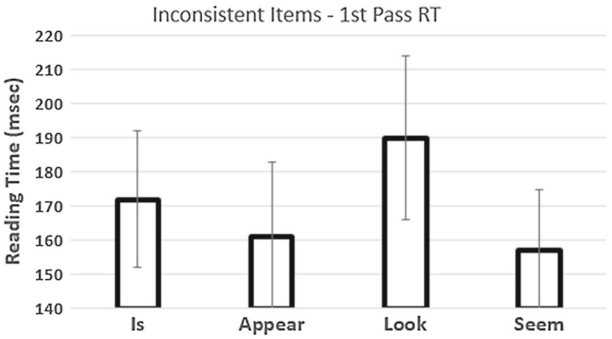

Fig. 5 Mean first pass reading times for the conflict region in inconsistent items. Error bars show the standard error of the mean

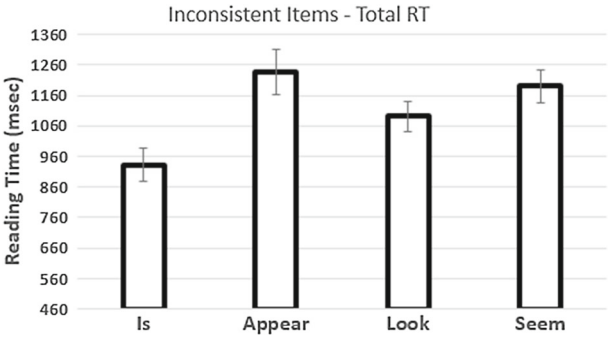

Fig. 6 Mean total reading times for the source region in inconsistent items. Error bars show the SE of the mean

The significant differences were between 'is' and the appearance verbs (appear-is: $t(46) = 5.02$, $p < 0.001$, look-is: $t(46) = -2.57$, $p < 0.05$, seem-is: $t(46) = -4.48$, $p < 0.001$). The comparison between 'look' and 'appear' was also significant $t(46) = 2.11$, $p < 0.05$, the differences between 'appear'-'seem' and 'look'-'seem' were not p's > 0.08. Observed differences were driven not by first-pass reading times (above)

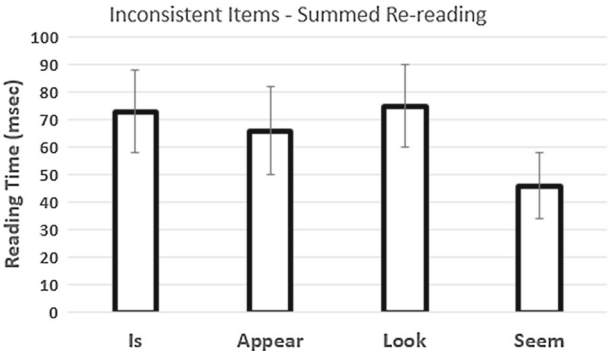

Fig. 7 Mean re-reading times for the conflict region in inconsistent items. Error bars show the SE of the mean

but by second-pass reading times that are indicative of integration difficulties (Sect. 4.3).

Second pass reading time: conflict region

Results showed no significant main effect of verb $F(3,138) = 1.28$, $p = 0.28$, $\eta^2 = 0.03$ (see Fig. 7). This is in line with first-pass reading times and the fact that while approximately 90% of trials showed a regression from the conflict region, only 22.2% of trials involved a return to the conflict region. This suggests that integration difficulties were addressed in re-reading the source region.

Appendix 3: Follow-up study on 'phenomenal' sense

We conducted an acceptability rating study to examine two hypotheses:

hI In ordinary discourse, all three appearance verbs have a 'phenomenal' sense or use that carries no doxastic implications (and allows us to say, e.g., that the round coin appears elliptical, when viewed sideways).

hII This sense or use is more salient for 'look' than for 'appear' and 'seem'.

Examination of hI is motivated by a reviewer query whether all three verbs have a phenomenal sense in ordinary discourse (or only in philosophy). Examination of hII is motivated by the mixed experimental picture concerning inappropriate doxastic inferences from 'look' (Sect. 4.5).

Methods

We recruited 51 participants with the same approach and from the same population as in Experiments 1–3 (Sect. 3). They were all native English-speakers (74.5% women, three non-binary, with an average age of 36.5, ranging from 16 to 77). Participants received instructions that explained with the help of an example that the same word

may be used in different senses, so that sentences that may appear to contradict each other can all be true, when using the word in different senses. They were told 'we are interested in whether the verbs "look", "appear", and "seem" are used in different senses that allow people to say different things that may appear to contradict each other'. They were instructed to 'take into account all senses of these verbs that you are familiar with', and asked to rate on a 7-point scale their confidence that given sentences using these verbs can be accepted as saying something true. Endpoints were explained as follows: '"1" means you are entirely confident that the sentence cannot be accepted as saying something true. "7" means you are entirely confident that the sentence can be accepted as saying something true. The mid-value "4" means you are really unsure whether or not the sentence can be accepted as saying something true.'

Participants rated 18 sentences that described familiar cases of non-veridical perception, where no adult perceiver is inclined to form a wrong belief about the shape, size, or colour of the object seen. Sentences used appearance-verbs either in a non-doxastic sense (to describe the looks of the object) or a doxastic sense (to state what property the viewer will think the object has), for example:

When you look at a round coin sideways, the coin looks elliptical. (non-doxastic, ND)
When you look at a round coin sideways, the coin looks round. (doxastic, D)

The first sentence is acceptable as true only if the verb has a non-doxastic reading (under the circumstances, nobody believes the coin is elliptical); the second is acceptable only if the verb has a purely doxastic reading (the viewer will believe the coin is round, but that is no description of what the coin 'phenomenally' looks like). Three pairs of sentences used shape adjectives, three used size, three used colour. Further examples include:

Seen from the beach, the huge ships anchored out at sea appear small/huge. (ND)/(D)
Under red light, white lab coats seem reddish/white. (ND)/(D)

Sentences were presented in random order. Verbs were rotated across items. Hypotheses concern non-doxastic (ND) items only. Doxastic items were intended as fillers, to ensure through contrast that participants would take into account the difference in adjectives (e.g., 'white'–'reddish') in the critical ND items. We thus manipulated a single variable (verb) with three levels. One participant was excluded as outlier (>3SDs from the mean in all conditions), prior to analysis.

Our analyses were driven by our a priori hypotheses. hI predicts that participants will accept ND items with all three verbs as saying something true (namely, in the non-doxastic sense), i.e., that acceptability ratings for ND items with all three verbs will be above neutral mid-point '4'. We assessed hI through three one-sample t-tests with a test value of 4. hII predicts higher ratings for ND items with 'look' than 'appear' or 'seem'. This results in a directional one-tailed hypothesis assessed via paired-samples t-tests.

Results and discussion

For ND items, mean ratings (with SDs) for 'look' (5.37, 1.09), 'appear' (5.03, 1.05), and 'seem' (4.97, 1.10) sentences were all significantly above mid-point 4 (appear: $t(49) = 6.92$, $p < 0.001$; look: $t(49) = 8.88$, $p < 0.001$; seem: $t(49) = 6.23$, $p < 0.001$). Paired-samples t-tests revealed that non-doxastic 'look' sentences attracted ratings significantly different from counterparts with 'appear' $t(49) = -1.70$, $p = 0.048$ and 'seem' $t(49) = 1.71$, $p = 0.047$. In contrast, ratings for ND sentences with 'appear' and 'seem' were not significantly different $t(49) = 0.31$, $p = 0.38$. For thoroughness, we also analysed results for doxastic items. Here, mean ratings (with SDs) for 'look' (3.12, 1.23), 'appear' (3.25, 1.25), and 'seem' (3.07, 1.09) sentences all were significantly below 4 (appear: $t(49) = -4.25$, $p < 0.001$; look: $t(49) = -5.05$, $p < 0.001$; seem: $t(49) = -6.00$, $p < 0.001$). Results for the paired-samples t-tests showed no significant differences between the conditions (all p's > 0.39).

Ratings for ND items were consistent with our hypotheses. However, since the doxastic sense or use of appearance verbs is well attested, in particular for 'appear' and 'seem' (Sect. 2.3), the ratings below neutral mid-point observed for doxastic (D) items with all three verbs suggest that participants answered a different question: How confident are you that the sentence is an acceptable (plausible or natural) thing to say about the situation envisaged? On this reading, the observed low ratings for D items are consistent with the dominance of the doxastic sense at any rate for 'appear' and 'seem': Participants did not find it plausible or natural to use this—familiar—sense to talk about familiar situations of non-veridical perception. This may be because the circumstances indicated do not change judgment (adult perceivers are too conversant with them), but do change the way things look, so that participants inferred that the sentence must be about the way things look. On this reading, high ratings for ND items suggest participants found it plausible or natural to invoke a non-doxastic phenomenal sense to talk about the way things look. But this implies they recognised, and were familiar with, such a sense. Also on this reading, observed ND ratings therefore support hI. That participants find it more natural to use this non-doxastic sense of 'look' than 'appear' or 'seem' suggests the sense is more prototypical for 'look' (see Fn7). This is consistent with higher salience, as per hII.

Appearance verbs arguably are the linguistic devices best suited to talk about the way things look in cases of non-veridical perception. The fact that, even so, the proportion of '6/7'-ratings remained overall low for ND items (53%) suggests that participants frequently found it impossible to completely suppress doxastic inferences from these verbs, even with strong contextual support (stronger than in the main study). The fact that the proportion of such ratings was lower for ND items with 'appear' (47%) and 'seem' (51%) than 'look' (56%) suggests that participants found it easier to completely suppress doxastic inferences from 'look'. This is predicted by the Salience Bias Hypothesis in conjunction with hII and the corollary that the doxastic sense is less dominant for 'look' than the other verbs. We conclude that all three appearance verbs have a non-doxastic 'phenomenal' sense in ordinary discourse, and this sense is more salient for 'look' than 'appear' and 'seem'.

Appendix 4: Discussion of H_2

The main study of the paper (reported in Sect. 4) also helps us assess the hypothesis H_2 as an alternative explanation of findings from a previous plausibility ranking experiment (Fischer et al. 2019). Instead of invoking inappropriate doxastic inferences, this explanation (Sect. 3.5) suggests that preferences for 'is' sentences in items like

The hill seemed/was quite steep. The rambler thought it was gentle

are based on how subjective versus objective the question under discussion (whether the hill is steep) is deemed to be: Participants prefer 'is' where this question is deemed objective, and appearance sentences where this is more subjective; the second sentences of items ('The rambler thought it was gentle') are not taken into account.

In the present main study, participants rated both items consistent and inconsistent with the doxastic inferences posited by H_1. This consistency manipulation in the second sentence affected plausibility ratings very strongly and more strongly than any other factor examined. This suggests that plausibility assessments also in the prior plausibility ranking study were influenced by the second sentences that affected consistency. Furthermore, in the present study, the consistency manipulation affected the plausibility of sentences with different verbs (slightly) differently, as we observed a (marginal) verb $\times$ consistency interaction. This suggests that where verbs have stronger stereotypical associations with doxastic patient properties, doxastic inferences reduced the plausibility of s-inconsistent sequels more strongly. We therefore infer that, in the earlier study, preferences were influenced by differences in verbs' strength of stereotypical association with doxastic inferences that were cancelled by the second sentence.

Crucially, the follow-up study (Sect. 4.4) showed that higher subjectivity ratings either failed to change plausibility ratings (in the s-consistent condition) or changed the plausibility of both appearance and 'is'-sentences in the same direction, and to pretty much the same extent (in the s-inconsistent condition). Differences in subjectivity ratings therefore cannot explain the observed differences in plausibility ratings between items with 'is' and with appearance verbs 'appear' and 'seem', in the main study, or the preferences observed in the earlier plausibility ranking study by Fischer et al. (2019) (with only s-inconsistent items).

Finally, H_2's alternative account requires that the patient role of verbs in appearance sentences be assigned to the author, rather than the protagonist, of the text (Sect. 3.1). This predicts a positive correlation between subjectivity and plausibility ratings for appearance sentences (Sect. 4.1). The findings of the follow-up study excludes this confound at any rate for items with visual objects. (The present study examined no others.) These findings speak against H_2 and suggest that, in the earlier study by Fischer et al. (2019), at any rate the preferences for 'is'-sentences over appearance sentences in items with visual objects were based on making and maintaining contextually inappropriate doxastic inferences about the protagonist.

Fischer and colleagues (2019) further observed attenuated preferences in items with abstract objects. They adduced this to a moderate amount of patient reassignment in such items, due to higher levels of perceived contradictoriness (Sect. 2.3). This explanation is inconsistent with the fact that, in our new Exp.1, s-inconsistent

appearance items with visual and with abstract objects were deemed equally contradictory, at any rate in items with verbs 'seem' and 'appear' (Sect. 3.2). The negative correlation between subjectivity and plausibility ratings observed in the follow-up study for s-inconsistent items with visual objects (Sect. 4.4) provide an alternative explanation for patient reassignment in items with abstract objects: Without patient reassignment, participants infer that protagonists have beliefs or opinions that clash with those attributed to them by the sequel; protagonists entertain clashing doxastic attitudes. The negative correlation means that participants find it less plausible that a protagonist has clashing doxastic attitudes towards matters of subjective opinion than towards matters of objective fact. This might reflect epistemic doubt: Participants take s-inconsistent items to convey that protagonists are second-guessing themselves. Second-guessing oneself may intuitively make more sense when the matter at hand is objective rather than subjective.

In Exp.3, most items with abstract objects were perceived as more subjective than most items with visual objects: Breaking down the 36 items by quartiles based on descending subjectivity rating, we observe: Of the nine most subjective items, all but one item had an abstract object (88.9%). Of the second nine items, all but two had an abstract object (77.7%). By contrast, of the least subjective nine items, all but one had visual objects and of the second to bottom nine all but two had visual objects. This means that 15 of 18 items with an abstract object (83.3%) were in the more subjective half of items, while 15 of 18 items with visual objects (83.3%) were in the less subjective half.

Even though appearance items with visual and abstract objects are perceived as equally contradictory, items with abstract objects tend to be perceived as more subjective, so that second-guessing oneself intuitively makes less sense and items are perceived as less plausible. Principles of charity have readers take remedial action to avoid gross implausibility. This extra source of implausibility may therefore push some appearance items with abstract objects over the threshold at which some participants resort to patient reassignment to obtain a more plausible interpretation of appearance sentences. This could account for the attenuation in preferences for 'is'-sentences observed by Fischer and colleagues (2019). We therefore take present findings to reinstate the findings from that previous study in the light of a potential confound.

References

Austin, J. L. (1962). *Sense and sensibilia*. Oxford: Oxford University Press.
Ayer, A. J. (1940). *Foundations of empirical knowledge*. London: Macmillan.
Ayer, A. J. (1956/1990). *The problem of knowledge*. London: Penguin.
Baz, A. (2017). *The crisis of method*. Oxford: OUP.
Beretta, A., Fiorentino, R., & Poeppel, D. (2005). The effects of homonymy and polysemy on lexical access: An MEG study. *Cognitive Brain Research, 24,* 57–65.
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.
Broad, C. D. (1923). *Scientific thought.* Repr. 2000. London: Routledge.
Brogaard, B. (2013). It's not what it seems: A semantic account of 'seems' and seemings. *Inquiry, 56,* 210–239.

Brogaard, B. (2014). The phenomenal use of 'look' and perceptual representation. *Philosophy Compass, 9*(7), 455–468.
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.
Clifton, C., Ferreira, F., Henderson, J. M., Inhoff, A. W., Liversedge, S. P., Reichle, E. D., et al. (2016). Eye movements in reading and information processing: Keith Rayner's 40 year legacy. *Journal of Memory and Language, 86,* 1–19.
Clifton, C., Staub, A., & Rayner, K. (2007). Eye movements in reading words and sentences. In R. P. G. van Gompel, et al. (Eds.), *Eye movements. A window on mind and brain* (pp. 341–371). Amsterdam: Elsevier.
Colaço, D., Kneer, M., Alexander, J., & Machery, E. (2018). *On second thought: A refutation of the reflection defense*. Pittsburgh: University of Pittsburgh. https://doi.org/10.13140/RG.2.2.34481.68967.
Crane, T., & French, C. (2015). The problem of perception. In N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Summer 2015. http://plato.stanford.edu/entries/perception-problem/.
Engelhardt, P. E., & Ferreira, F. (2016). Reaching sentence and reference meaning. In P. Knoeferle, P. Pyykkonen, & M. W. Crocker (Eds.), *Visually situated language comprehension* (pp. 127–150). Amsterdam: John Benjamins.
Fein, O., Yeari, M., & Giora, R. (2015). On the priority of salience-based interpretations: The case of sarcastic irony. *Intercultural Pragmatics, 12,* 1–32.
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). Verbal fallacies and philosophical intuitions: The continuing relevance of ordinary language analysis. In B. Garvey (Ed.), *J.L. Austin on language* (pp. 124–140). Basingstoke: Palgrave.
Fischer, E., & Curtis, M. (Eds.). (2019). *Methodological advances in experimental philosophy*. London: Bloomsbury.
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. (2019a). Lingering stereotypes: Salience bias in philosophical argument. *Mind and Language, 2019,* 1–25. https://doi.org/10.1111/mila.12249.
Fischer, E., & Engelhardt, P. E. (2019b). Eyes as windows to minds: Psycholinguistics for experimental philosophy. In E. Fischer & M. Curtis (Eds.), *Methodological advances in experimental philosophy* (pp. 43–100). London: Bloomsbury.
Fischer, E., Engelhardt, P. E., & Herbelot, A. (2015). Intuitions and illusions: From experiment and explanation to assessment. In E. Fischer & J. Collins (Eds.), *Experimental philosophy, rationalism and naturalism* (pp. 259–292). London: Routledge.
Fischer, E., Engelhardt, P. E., Horvath, J., & Ohtani, H. (2019). Experimental ordinary language philosophy: A cross-linguistic study of defeasible default inferences. *Synthese*. https://doi.org/10.1007/s11229-019-02081-4.
Fischer, E., & Sytsma, J. (2020). *Zombie intuitions*. University of East Anglia, Ms.
Fish, W. (2010). *Philosophy of Perception*. London: Routledge.
Garrett, M., & Harnish, R. M. (2007). Experimental pragmatics: Testing for implicatures. *Pragmatics & Cognition, 17,* 245–262.
Giora, R. (2003). *On our mind. Salience, context, and figurative language*. Oxford: OUP.
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.
Goldberg, A. E. (2003). Constructions: A new theoretical approach to language. *Trends in Cognitive Sciences, 7,* 219–224.
Grice, H. P. (1961). The causal theory of perception. *Proceedings of the Aristotelian Society, 35,* 121–152.

Grice, H. P. (1989). Logic and conversation. In H. P. Grice (Ed.), *Studies in the ways of words* (pp. 22–40). Cambridge, MA: Harvard UP.

Guterstam, A., Keana, H. H., Webba, T. W., Keana, F. S., & Grazianoa, M. S. A. (2019). Implicit model of other people's visual attention as an invisible, force-carrying beam projecting from the eyes. *Proceedings of the National Academy of Sciences, 116,* 328–333.

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. (2018). 'Nobody would really talk that way!': The critical project in contemporary ordinary language philosophy. *Synthese*. https://doi.org/10.1007/s11229-018-1812-x.

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. A representative theory*. Cambridge: CUP.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, et al. (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 49–81). 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.

Klepousniotou, E., Pike, B., Steinhauer, K., & Gracco, V. (2012). Not all ambiguous words are created equal: An EEG investigation of homonymy and polysemy. *Brain and Language, 123,* 11–21.

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.

Leech, G., Payson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English: Based on the British National Corpus*. London: Longman.

Levinson, S. C. (2000). *Presumptive meanings. The theory of generalized conversational implicature*. Cambridge: MIT Press.

Livengood, J., & Sytsma, J. (2020). Actual causation and compositionality. *Philosophy of Science, 87,* 43–69.

Livengood, J., Sytsma, J., & Rose, D. (2017). Following the FAD: Folk attributions and theories of actual causation. *Review of Philosophy and Psychology, 8,* 274–294.

MacGregor, L. J., Bouwsema, J., & Klepousniotou, E. (2015). Sustained meaning activation for polysemous but not homonymous words: Evidence from EEG. *Neuropsychologia, 68,* 126–138.

Machery, E. (2017). *Philosophy within its proper bounds*. 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.

Maund, J. B. (1986). The phenomenal and other uses of 'looks'. *Australasian Journal of Philosophy, 64,* 170–180.

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.

Moore, G. E. (1918/19). Some judgments of perception. *Proceedings of the Aristotelian Society*, *19,* 1–29.

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

Nichols, S., & Pinillos, N. Á. (2018). Skepticism and the acquisition of "knowledge". *Mind and Language, 33,* 397–414.

Price, H. H. (1932). *Perception* (2nd edn.), repr. 1961. London: Methuen.

Pylkkänen, L., Llinás, R., & Murphy, G. L. (2006). The representation of polysemy: MEG evidence. *Journal of Cognitive Neuroscience, 18,* 97–109.

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

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*, 372–422.
Rayner, K., Warren, T., Juhasz, B. J., & Liversedge, S. P. (2004). The effect of plausibility on eye movements in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 1290–1301.
Robinson, H. (1994). *Perception*. London: Routledge.
Russell, B. (1912/1980). *The problems of philosophy*. Oxford: OUP.
Saint-Germier, P. (2019). Getting gettier straight: Thought experiments, deviant realization, and pragmatic enrichment. *Synthese*. https://doi.org/10.1007/s11229-019-02166-0.
Schwitzgebel, E., & Ellis, J. (2017). Rationalization in moral and philosophical thought. In J.-F. Bonnefon & B. Trémolière (Eds.), *Moral inferences* (pp. 170–190). Lomdon: Psychology Press.
Shtulman, A., & Valcarcel, J. (2012). Scientific knowledge suppresses but does not supplant earlier intuitions. *Cognition, 124,* 209–215.
Sirois, S., & Brisson, J. (2014). Pupillometry. *WIREs Cognitive Science, 5,* 679–692.
Smith, A. D. (2002). *The problem of perception*. Cambridge, MA: Harvard UP.
Snowdon, P. F. (1992). How to interpret 'direct perception'. In T. Crane (Ed.), *The contents of experience*. Cambridge: CUP.
Snowdon, P. F. (2015). Sense-data. In M. Matthen (Ed.), *The Oxford handbook of philosophy of perception* (pp. 118–135). Oxford: OUP.
Sytsma, J. (2019). *Objectivity, not salience bias? Commentary on Fischer et al. (2019).* The Brains Blog. http://philosophyofbrains.com/2019/07/15/symposium-on-fischer-etal-experimental-ordinary-language-philosophy.aspx.
Sytsma, J., & Livengood, J. (2016). *The theory and practice of experimental philosophy* **(Broadview)**.
Thagard, P., & Nisbett, R. E. (1983). Rationality and charity. *Philosophy of Science, 50,* 250–267.
Thompson, V., & Evans, J. St. B. T. (2012). Belief bias in informal reasoning. *Thinking & Reasoning, 18,* 278–310.
Weinberg, J. M. (2017). What is negative experimental philosophy good for? In G. D'Oro & S. Overgaard (Eds.), *The Cambridge companion to philosophical methodology* (pp. 161–183). Cambridge: CUP.
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.