The experimental relaunch of critical ordinary language philosophy
Response to Comments

Eugen Fischer and Paul Engelhardt (UEA)

In 'Experimental ordinary language philosophy: a cross-linguistic investigation of default inferences' (Fischer et al., 2019), we explored with Joachim Horvath and Hiroshi Ohtani how experimental methods and findings from psycholinguistics can be adapted to analyse philosophical arguments and to relaunch critical ordinary language philosophy in the wake of J.L. Austin (1962). The Austinian research program addresses philosophical problems that are generated by philosophical paradoxes. It seeks to 'dissolve' such problems by exposing seductive verbal fallacies at the root of the relevant paradoxical arguments. The fallacies Austin focuses on are contextually inappropriate default inferences that are routinely made in language comprehension and production. Our paper addresses the crucial methodological challenge to this approach: Austin maintains – and psycholinguistic research confirms – that competent speakers' inferences are highly sensitive to contextual cues. So why should philosophical proponents of the relevant arguments – who are competent speakers – make the contextually inappropriate inferences supposedly involved in their arguments?

More specifically, our paper sought to do three things: First, we developed and experimentally tested a psycholinguistic explanation that lets us understand when and why even competent speakers make contextually inappropriate stereotypical inferences from polysemous words. We developed the Salience Bias Hypothesis that rare uses of such words trigger inappropriate inferences which are licensed only by the words' dominant sense, when this sense is clearly dominant and functional for the interpretation of less salient (rare) uses. We tested this hypothesis through experiments on inappropriate doxastic inferences from rare phenomenal uses of appearance verbs. Second, we attempted to show that our findings help expose fallacies in 'arguments from illusion' and help 'dissolve' the 'problem of perception' (Crane & French, 2015) that is generated by these arguments together with structurally similar 'arguments from hallucination' (analysed elsewhere, see Fischer & Engelhardt, 2019a; 2019b). Third, we explored how insights from cross-cultural psycholinguistics could be deployed to assess the philosophical relevance of the inappropriate inferences we experimentally documented.

In this symposium, two sets of acute and helpful comments examine our efforts towards the first and second of these aims, respectively: Justin Sytsma assesses our experimental evidence, proposes an alternative explanation of our findings, and presents some empirical data to adjudicate between our explanation and the alternative. Pendaran Roberts, Keith Allen and Kelly Schmidtke critically assess the relevance of our empirical findings to the analysis of the argument from illusion. This response will defend the account of the target paper against these important challenges, by drawing on further experimental work we undertook in the meantime. Section 1 will outline a new eye tracking study (to be properly reported elsewhere). Section 2 will deploy the findings to assess Sytsma's objections. Section 3 will examine his new data and alternative explanation. Section 4 will discuss the relevance of our findings to the analysis of arguments from illusion, in response to Roberts, Allen and Schmidtke. This exchange might usefully illustrate what Sytsma aptly called the 'difficulty of doing experimental philosophy' and demonstrate how we can address some of the relevant difficulties by expanding the methodological repertoire of experimental philosophy beyond questionnaire-based methods.

1. An eye-tracking study

Psycholinguists use a 'cancellation paradigm' to study automatic comprehension inferences from words: Participants read sentences or short texts in which the word of interest is followed by a sequel that is inconsistent with the hypothesised inference from the word. E.g., to test for inferences from an appearance verb ('X seems F [to S]') to doxastic conclusions (*S thinks that X is F*), this paradigm has us use sentences like

(1) The dress seemed blue. Hannah thought it was green.

If the hypothesised inference is made, the clash of the conclusion (*Hannah thought the dress was blue*) with the sequel ('Hannah thought it was green') will engender comprehension difficulties requiring cognitive effort. This effort is picked up by a variety of measures including pupil dilations (Sirois & Brisson, 2014), longer reading times (Clifton et al., 2007), and signature electrophysiological responses ('N400s') (Kutas & Federmeier, 2011). The inference may be swiftly suppressed and fail to influence further judgment and reasoning (Fischer & Engelhardt, 2017) – in which case it presumably is of no philosophical interest. If it is not suppressed, however, participants will find sentences like (1) less plausible than counterparts that use a contrast verb that triggers no such inference (like 'was' in 2 below) or have a sequel that is consistent with the relevant inference (as in 3 below).

(2) The dress was blue. Hannah thought it was green.
(3) The dress seemed blue. Hannah thought it was navy.

In the target paper, we used a simple questionnaire-based approach that implemented the cancellation paradigm with a speeded forced-choice plausibility ranking task, in English, German, and Japanese: We paired appearance-sentences like (1) with is-sentences like (2), and participants judged which of the two struck them as more plausible. 'Is'-sentences are mildly implausible, insofar as they claim that the viewer got quite obvious things wrong. If appearance verbs ('appear', 'seem', 'look') trigger doxastic inferences and are understood to take, e.g., Hannah as patient (i.e., the dress seemed blue to Hannah), then appearance-sentences like (1) are yet more implausible, as they engender a contradiction. Participants will then judge 'is'-sentences (like 2) more plausible than appearance-counterparts (like 1). In ordinary discourse, appearance verbs are used mainly to attribute beliefs to often implicit patients. In the argument from illusion, these verbs are intended in a phenomenal sense, in which they merely describe the subjective experience of a subject and say nothing about what the subject believes. In our experiment, the sequel (e.g. 'Hannah thought it was green') renders a phenomenal reinterpretation of the appearance verb contextually appropriate. Consistent preferences of 'is'-sentences over alternatives therefore provide first evidence that contextually inappropriate doxastic inferences are made and maintained.

In a follow-up study, we implemented the cancellation paradigm with eye tracking: We used reading-time measurements to garner evidence of automatic inferences and employed plausibility ratings to assess their influence on subsequent cognition. When we read sentences, our eyes may pass over the same words several times. Whereas first pass reading times are largely determined by word length, word frequency, and the word's predictability in context, difficulties in integrating information from different parts of the sentence may have us reread bits of the sentence. In particular, 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 will therefore lead to higher 'late' (second-pass and total) reading times for either the conflict region ('green') or the source region ('seemed blue'), or both. We therefore manipulated both consistency (e.g. (1) vs (3) above) and verb ('appear', 'seem', 'look', 'is'). In this setting, the hypothesis that appearance verbs trigger doxastic inferences predicts, among other things, higher late reading times for conflict or source regions in inconsistent appearance-items (like 1 above) than in 'is'-counterparts (like 2 above). The further hypothesis that these doxastic inferences do not get swiftly suppressed but influence further cognition predicts that, in a subsequent plausibility rating task, inconsistent appearance items (like 1 above) will be deemed less plausible than inconsistent 'is'-items (like 2 above) and consistent appearance items (like 3 above).

In our eye-tracking study, 48 psychology undergraduates from the University of East Anglia, all native speakers of English, read 48 critical items (six per condition) and an equal number of fillers. To prevent floor effects, a prior norming study excluded inconsistent items where 'is'-versions were rated below 2.5 on a 5-point Likert scale. We manipulated consistency and verb within subjects, in a 2×4 design. Items were rotated across lists. On each trial, an item was presented in a single line on a computer screen, superseded by a 5-point plausibility rating prompt.

For plausibility ratings, we observed the predicted main effect of consistency $F(1,46) = 387.21$, $p < .001$ and verb $F(1,46) = 7.53$, $p < .01$. In addition, there was a marginal 2-way interaction $F(1,46) = 3.29$, $p = .076$ (see Figure 1). Participants rated consistent items distinctly plausible, or significantly above neutral mid-point (all *p*'s < .001), and deemed items with different verbs equally plausible $F(3,138) = .59$, $p = .62$. Inconsistent items with all verbs were deemed distinctly implausible, or significantly below mid-point (all *p*'s < .001), and there were significant differences between verb conditions $F(3,138) = 3.54$, $p < .05$. As predicted, items with 'appear' and 'seem' were deemed less plausible than items with the contrast verb 'is' (appear vs. is: $t(46) = -2.09$, $p < .05$; seem vs. is: $t(46) = 2.65$, $p < .05$). The mean plausibility rating for items with 'look' was numerically lower than the mean rating for 'is'-items, but, against predictions, this difference was not significant $t(46) = .36$, $p = .72$.

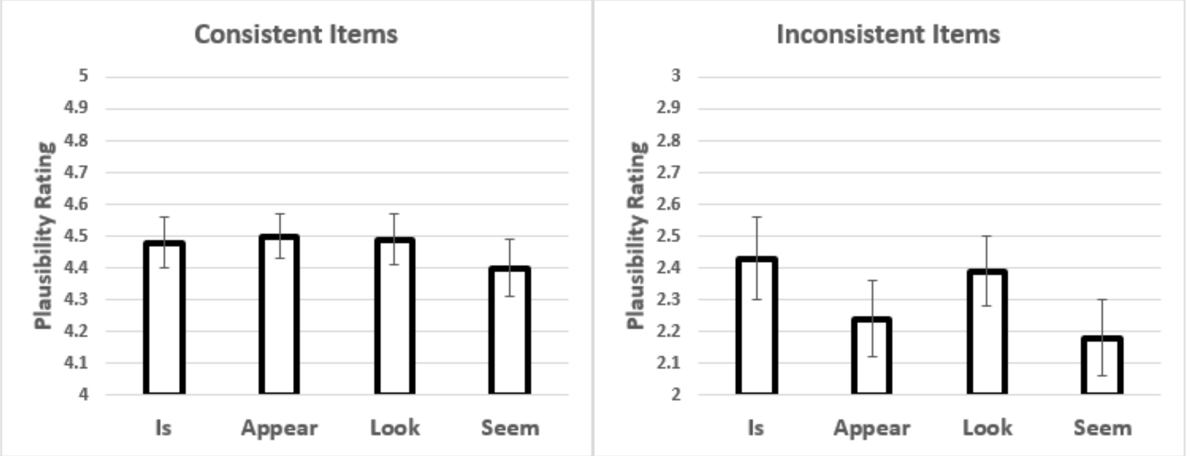

Figure 1. Mean plausibility ratings. Error bars show the standard error of the mean.

For eye movements, results were partially consistent with our predictions – and partially unexpected. In a previous study that used the same paradigm to document automatic inferences from perception verbs (Fischer & Engelhardt, 2019a), we had found higher late reading times in the conflict region for items with critical verbs than with contrast verbs. In the present study, by contrast, all the extra processing time was devoted to the source region. Given the patterns in plausibility ratings and the predictions, we focused the eye movement analysis solely on the inconsistent items. Total reading times for the conflict region (e.g., 'green' in 1 above) were not significantly different for items with the contrast verb 'is' than for the appearance verbs *F*(3,138) = 1.39, *p* = .25 (see Figure 2). Second-pass re-reading times (defined as total reading time minus first-pass reading times) were summed across the first verb and first object (i.e. the source region, e.g., 'seems blue' in 1 above). Results showed a significant effect of verb *F*(3,138) = 4.65, *p* < .01. Paired comparisons revealed summed re-reading times were appreciably higher in 'appear'-items than 'is'-items *t*(46) = 3.41, *p* < .01 and in 'seem'-items than in 'is'-items *t*(46) = -2.85, *p* < .01. The difference between 'look'-items and 'is'-items remained shy of significance *t*(46) -1.34, *p* = .19. For purposes of subsequent discussion, we further note that differences were marginally significant between 'look'- and 'appear'-items *t*(46) 1.92, *p* = .06, though not between 'look'- and 'seem'-items *t*(46) = -1.55, *p* = .13.

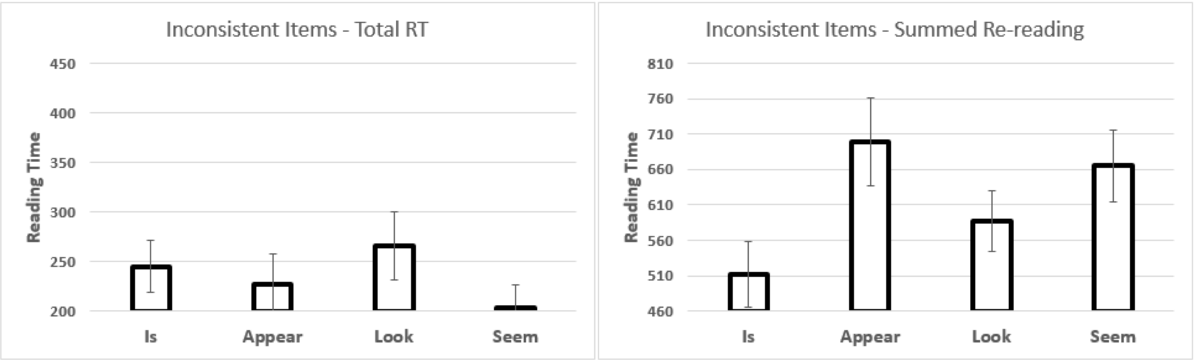

Figure 2. Left panel shows total reading time on the **conflict region**. Right panel shows the summed re-reading time on verb and adjective jointly making up the **source region**. Error bars show the standard error of the mean.

These findings provide evidence of automatic doxastic inferences at any rate from 'appear' and 'seem' (the appearance verbs most frequently used in arguments from illusion) and suggest that these inferences go on to influence further judgment and reasoning. The findings also give rise to some exciting questions, e.g.: Why did all the extra processing action take place on the source-, rather than the conflict region? And why did 'appear' and 'seem', but not 'look', trigger doxastic inferences that influence further cognition? The second question may be answered by our Salience Bias Hypothesis: One of the factors determining whether contextually inappropriate inferences go through unsuppressed is just how dominant the sense supporting these inferences is. Evidence from distributional semantic analysis suggests that while all appearance verbs have a doxastic use (to attribute beliefs to patients), this is clearly dominant in the case of 'seem' and 'appear', whereas its dominance is less pronounced for 'look' (Fischer, Engelhardt, & Herbelot, 2015). This difference in salience might explain the processing differences between the verbs, which we observe now. However, in two earlier studies which used the forced-choice plausibility ranking paradigm (Fischer & Engelhardt, 2016; and the present target paper, Fischer et al., 2019), 'look'-items had performed no different from 'appear' items. The question remains for another occasion. In the remainder of this paper, we will explore how the new findings help us address the acute objections of the commentators.

2. Sytsma's objections

Justin Sytsma challenges our experimental evidence on three counts, before assessing it in the light of new data and an alternative explanation of our findings. The forced-choice plausibility ranking task in the target paper was speeded. Sytsma's first worry is that our findings might therefore not warrant conclusions about thinkers who are under no time constraints. In response, we observe that in our follow-up study plausibility ratings were not speeded but self-paced, and participants' ratings, even so, largely mirrored prior rankings: Inconsistent 'is'-items were rated more plausible than appearance counterparts, and this difference was significant for items with 'appear' and 'seem'.

We read Sytsma's second worry as an instance of the 'expertise objection' to experimental philosophy (*cf.* Machery, 2017, 158-169): Findings about lay participants need not warrant conclusions about expert philosophers (like proponents of the argument from illusion). To assess the extent to which this important objection applies to our studies, we may consider trial order effects: The most relevant advantage expert philosophers should have over our undergraduate participants is that analytic philosophers have reflected on differences in meaning and use between related words or senses, so that their judgments may have become more sensitive to such differences. Philosophers might therefore be more sensitive than lay participants to the differences in meaning and use between different, but related appearance verbs, or might be more inclined to invoke the phenomenal sense of these verbs, where their ordinary doxastic implications are cancelled. These 'philosophical advantages' should to some extent be mimicked by learning effects occurring over the course of our experiments, which expose participants to similar items with subtle manipulations. Continued exposure to similar items with different verbs might make participants' judgments more sensitive to differences in meaning and use between those different verbs. Alternatively, continued exposure to the consistency manipulation in our follow-up study might make participants more sensitive to the availability of doxastic *and* phenomenal interpretations of appearance verbs. Increased sensitivity of either kind might show up in performance differences between the first and the second half of trials.

In our follow-up study, we indeed found a pronounced trial order effect: In the first half of trials, plausibility judgments did not significantly differ between items with different verbs in the inconsistent condition (appear vs. is: $t(44) = -1.2$, $p = .24$; seem vs. is: $t(42) = -1.61$, $p = .11$; all other p's > .38). The observed sensitivity to the verb manipulation arose from performance on the second half of trials, suggesting that our participants became more sensitive to the difference in doxastic implications between 'seem' and 'appear' vs. 'is'. Increased sensitivity to the availability of the phenomenal interpretation of appearance verbs would have had the opposite effect, as the phenomenal interpretation renders 'inconsistent' items with appearance verbs consistent and more plausible.

Reanalysing the data from the previous forced-choice plausibility ranking study, we found evidence of the same trial order effect: In the English experiment in this study (Fischer et al., 2019, Exp.1), the predicted tendency to prefer 'is'-sentences with visual objects over counterparts with appearance verbs significantly increased between the first and the second half of trials (look .57 vs. .79, $t(72) = -4.85$, $p = .000$, $\eta^2 = .25$; appear .59 vs. .73, $t(72) = -3.05$, $p = .003$, $\eta^2 = .11$; seem .53 vs. .66, $t(72) = -2.98$, $p = .004$, $\eta^2 = .11$). The participants in this experiment were psychology undergraduates. By contrast, the participants in the German experiment (Fischer et al., 2019, Exp.2) were philosophy students who had already benefited from at least half a semester of philosophical training. If trial order effects in our studies mimic the advantages of the philosophically trained over laypeople, 'is'-preferences in the German study should be higher from the start. This is what we found in reanalysing our German data, where 'is'-preferences in the first half of trials (look .71; appear .74; seem .74) were similar to English psychology students' preferences in the second half of trials (above), and performance differences between first and second half of trials were less pronounced: still significant for 'appear' ($t(47) = -.2.72$, $p = .009$, $\eta^2 = .14$), marginally so for 'seem' ($t(47) = 1.91$, $p = .063$, $\eta^2 = .07$), and non-significant for 'look' ($t(47) = -.903$, $p = .37$, $\eta^2 = .02$).

To sum up: As exposure to our subtly manipulated items rendered our participants more similar to expert philosophers, initial doxastic inferences began to influence subsequent cognition *more* strongly, while there was no evidence of increased alignment of judgments with contextually appropriate 'phenomenal reinterpretation'. The performance of participants without philosophical training increased significantly and became similar to the initial performance of participants with some philosophical training, who displayed less pronounced learning effects. These findings are suggestive. While they do not obviate the need for further research with expert philosophers, they do provide initial support for the hypothesis that also philosophers will be inclined to make contextually inappropriate doxastic inferences from phenomenal uses of appearance verbs, which influence further judgment.

The third of Sytsma's initial worries is that while our forced-choice ranking task yielded the significant preferences for 'is'-counterparts over inconsistent appearance-sentences that we predicted, the preference was not particularly pronounced, at 64.3% across all relevant English items. (It was 73.3% across all relevant German items, and 60.3% for Japanese.) Our follow-up study replicated this subtle effect, where inconsistent appearance-sentences were deemed less plausible than is-counterparts, but all items were deemed distinctly implausible, and the differences between the means, while mostly significant, where not pronounced (Sec. 1). Our follow-up findings are, however, more robust by complementing outcome with process measures. The latter provide evidence of doxastic inferences in the shape of significant and pronounced differences in total and second pass reading times for the source region.

We think the subtlety of the – albeit significant – difference in plausibility rankings and ratings is due to a difficulty that arises when we use the cancellation paradigm to study doxastic inferences from appearance verbs. As explained above (Sec. 1), the clash between these inferences and the sequel is not the only source of implausibility in inconsistent items: Rather, also 'is'-counterparts ask us to accept that the protagonist got things wrong; and since the issues at issue (blue or green colour of a dress, steep or gentle slope of a hill) seem rather basic, this strikes our participants as distinctly implausible. Despite the norming study designed to prevent floor effects (see Sec.1), the doxastic inferences in inconsistent appearance items can hence merely make a limited difference.[1]

The scope for divergence in plausibility between appearance- and contrast-items is further reduced by the fact that participants who keenly feel the conflict between doxastic inferences and inconsistent sequels have the option of avoiding a contradictory interpretation by reassigning the patient role of the appearance verb, from the text's protagonist (e.g. Hannah) to its author (the dress seems blue not to Hannah but to the author). This turns the first sentence into the expression of a hedged judgment about, e.g., the dress (*I think the dress is blue*), which the author could have expressed more simply by writing, e.g., 'The dress is

[1] Unhelpfully, in the main study participants rated 'inconsistent' is-items even lower than in the norming study.

blue'. Preference of an appearance-verb over the simpler 'is' implies (with the Maxim of Manner) that doubt-and-denial conditions obtain (Grice, 1961). These conditions make it more plausible that the protagonist (Hannah) should have a different belief; i.e., competing pragmatic inferences that are higher in the pragmatic pecking order may defeat the stereotypical inferences of interest (*cf.* Levinson, 2000, pp.157-158). In our experimental paradigm, participants will then find the appearance-sentence more plausible than the 'is'-counterpart. Indeed, the target paper employed the otherwise crude plausibility ranking task (which requires participants to compare appearance-sentences with otherwise identical 'is'-sentences) precisely to invite competing Manner-inferences and examine to what extent they defeat the inferences of interest.

The upshot is that neither of the implementations of the cancellation paradigm we used leaves scope for dramatic differences in preferences or plausibility ratings between verb conditions. The preferences we observe in the plausibility ranking study with German philosophy students (73.3% for 'is'; Fischer et al., 2019; Exp.2) and the different plausibility ratings we observed in our eye tracking study with English psychology undergraduates during the second half of trials (appear 2.1, seem 2.16, is 2.37) are probably the best evidence we can obtain for our hypothesis with the experimental design and samples we used in these studies. This evidence reflects an experimental design that only leaves the inferences of interest limited scope to make a difference to the judgments measured. It does not warrant the conclusion that the inferences documented through reading times influence further cognition only weakly.[2]

3. Sytsma's experiments and alternative explanation

Our target paper sought to provide evidence for our key hypothesis that contextually inappropriate doxastic inferences from phenomenal uses of appearance verb are made – and the further hypothesis that these inferences are not defeated by competing pragmatic inferences. The above reasoning (Sec. 2) had us infer that such defeat is the more likely the more contradictory items seem to participants. We assumed 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 'The hill seemed quite steep. The rambler thought it was gentle'). We accordingly hypothesised that pragmatic defeat would occur in items with abstract objects but not in items with visual objects. This hypothesis predicted attenuated preferences for 'is'-sentences in items with abstract, rather than visual objects. By confirming this prediction, we hoped to garner compelling evidence that the stereotypical inferences of interest are not defeated in the perceptual cases of interest: The attenuation of 'is'-preferences concerning items with abstract objects would show that our 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.

Sytsma takes this comparison between items with visual and abstract objects as point of departure for an ingenious line of criticism that calls into question our experimental support for our key hypothesis that inappropriate doxastic inferences are made when participants read items with visual objects. He suggests that these items are perceived as no less contradictory than items with abstract objects. Therefore, he argues, patient reassignment should occur as

[2] However, the experimental design also makes it impossible to generate evidence that the inferences of interest influence further cognition strongly – which was Sytsma's main intended objection (personal communication).

frequently in items with visual objects, and the preferences for is-sentences over appearance-sentences that we observed in such items cannot be due to absence vs. presence of inferences to belief attributions (*The rambler believed the hill was quite steep*) which clash with the sequel ('The rambler thought the hill was gentle'). Rather, our participants reassigned the patient role of appearance verbs to us, the authors, and read appearance sentences as hedged authorial judgments. This makes our appearance-sentences more appropriate expressions of subjective opinions and our 'is'-sentences more appropriate expressions of uncontroversial objective facts. Sytsma suggests that our participants preferred 'is'-sentences in items with visual objects because they regarded them as stating objective facts and had a less clear-cut preference when considering items with abstract objects, because they understood these items to express a more subjective opinion (rendering appearance-sentences more appropriate). Sytsma helpfully conducts three experiments to assess this alternative explanation of our data.

His first study was intended to assess our assumption that items with abstract objects would be perceived as more contradictory than our items with visual objects. Sytsma's items made explicit the doxastic inference we claimed to be triggered by the first sentences in our items (e.g., 'The rambler thought the hill was quite steep. The rambler thought the hill was quite gentle'). Participants rated how contradictory these items are. Consistent with our assumption, participants regarded items with abstract objects as significantly more contradictory than items with visual objects. However, the effect size remained just shy of medium, and the mean ratings for both visual and abstract items were significantly above neutral mid-point. This led Sytsma to conclude that it remains unclear whether the differences observed are large enough to explain our original findings.

We welcome this study, but observe it does not fully engage with our reasoning. Our thought was this: In conjunction with visual objects, appearance verbs will trigger doxastic, epistemic, and experiential inferences supported – with decreasing strength – by the associated situation schema. Readers will infer that the rambler thought the hill was quite steep; that he presumably knew this; that he was possibly looking at the hill, which looked steep to him, etc. Abstract objects render these last inferences (*Cole was looking at the plan*) contextually inappropriate, and they are swiftly suppressed. In abstract items, the first sentence therefore boils down to the attribution of a belief, and the conflict with the sequel ('Cole believed it was terrible') is highly salient. In visual items, by contrast, the sequel clashes only with some of the maintained inferences, and these items do not reduce to a mere contradiction. To test our assumption, it might therefore have been better to elicit contradictoriness ratings for our original items, or variants which make explicit the intended assignment of patient role ('The hill seemed quite steep to the rambler. The rambler thought the hill was gentle').

Sytsma's second study sought to test our hypothesis that, at any rate for visual items, readers will assign the patient-role of the appearance verb in the first sentence to the agent of the second sentence (e.g., the rambler). Participants read our initial items and were asked who they took to make the claim expressed by the first sentence, e.g., 'The hill seemed quite steep' – 'the rambler' or 'the author'. Participants chose 'the author' in most cases, and no less frequently for items with visual objects than with abstract objects. However, we do not think that our hypothesis can be examined with this transparent design. Once the contradiction-avoiding assignment (of the patient-role to the author) is suggested to them, participants will of course opt for it. But that does not show that they would have come up with it, otherwise. An example of such task artifactuality can be provided with the help of this vignette:

> 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?

Many readers find it difficult to work this out when first encountering the vignette. One explanation is that they automatically infer that the speaker has the gender stereotypically associated with 'surgeon' and do not manage to come up with a situation model in which the surgeon is the young man's mother. This explanation could not be refuted by asking participants, 'Is the surgeon the father of the young man or the mother?', and noting that they overwhelmingly respond, 'the mother', once the question has suggested this possibility to them.

Our follow-up study addresses the issue of patient-role assignment by eliciting plausibility ratings for both inconsistent and consistent items with visual objects (like 1 and 3 above, respectively). If the patient-role is assigned to the author when participants read consistent items, readers will infer that doubt-and-denial conditions obtain. This will make it less plausible that, e.g., Hannah should think the dress is navy, when the dress 'seems blue', than when it 'is blue'. The finding that plausibility ratings for consistent items were the same in all verb conditions therefore suggests that at any rate for such items, readers assigned the patient-role to the text's protagonist (Hannah), rather than the author. Lower plausibility ratings for appearance- than 'is'-versions of inconsistent items provide evidence that readers frequently do not reassign the patient-role from protagonist to author, when coming across an inconsistent sequel, either (above, Sec.2). However, also this follow-up study does not allow us to determine more precisely to what extent readers then continue to assign the patient role to protagonists, and further research is required.

Sytsma's third experiment examined his alternative explanation of our data: Our participants preferred 'is'-sentences over appearance-counterparts in items with visual objects, and did so more strongly in such items than in items with abstract objects. But while the appearance verbs do trigger a doxastic inference, inconsistent sequels prompt reassignment of the patient-role from the protagonist to the author, and as a result the first sentence ('The hill seemed quite steep') is read as expression of the author's subjective opinion, whereas the 'is'-counterpart is read as expressing an objective fact. Sytsma then hypothesises that our participants regarded our items' claims about visual objects as objective, and their claims about abstract objects as more subjective. This would account for the preferences of 'is'-sentences in items with visual objects, and the attenuated preferences in items with abstract objects. To examine his hypothesis, Sytsma elicited ratings of subjectivity vs. objectivity for the statements expressed by the 'is'-sentences used in our items. Findings confirmed the hypothesis.

We do not think that our initial plausibility ranking study on its own can address the potential confound thus established. Our follow-up study, however, may speak to it: Pertinent evidence emerges when we interpret re-reading times in conjunction with plausibility ratings. In line with his account, Sytsma (personal communication) interprets increased rereading times for the source region as evidence of patient reassignment to the appearance verb. For the inconsistent items in our new study, his account then predicts patient reassignment for all appearance verbs ('seemed blue'), resulting in higher rereading times for source regions of appearance items than for ‘is’-items; furthermore, it predicts no differences between different appearance verbs. Our study, however, observed no significant differences in rereading times between inconsistent ‘look’- and ‘is’-items; but it did observe a marginally significant difference between ‘look’- and ‘appear’-items (Sec.2).

Moreover, if, as a result of patient reassignment, participants read the first sentence as expression of the author’s opinion, inconsistent appearance items should strike them as no less plausible than inconsistent ‘is’-items: The ‘is’-items state a fact (e.g., the dress was blue) and then suggest that the protagonist held a wrong belief about it (Hannah thought it was green). The appearance items, on Sytsma’s account, express the subjective opinion of someone else (the author). But that people hold different opinions is no less plausible than that the protagonist should get it wrong. (Put differently, that at least one of two people should get it wrong is no less plausible than that one person should get it wrong.) However, our study observed significantly lower plausibility ratings for ‘appear’- and ‘seem’-items than ‘is’-items. Sytsma’s account does not seem able to explain this finding – or why ‘look’-items are deemed as plausible as ‘is’-items.

Our account, by contrast, can build on the finding that ‘look’ has a less strong stereotypical association with doxastic patient properties than ‘appear’ and ‘seem’, as evidenced by distributional semantic analysis (Fischer, Engelhardt, & Herbelot, 2015). Conclusions of initial doxastic inferences (*Hannah thinks that the dress is blue*) are therefore easier to suppress in the face of conflict with the sequel, when triggered by ‘look’ than by ‘appear’ or ‘seem’, and the lesser suppression effort with ‘look’ leads to greater suppression success. We interpret rereading times for the source region as evidence of such suppression effort – and indeed observed that rereading times for ‘look’-items were not significantly higher than for ‘is’-items, while rereading times for ‘appear’- and ‘seem’-items were appreciably higher than for ‘is’-counterparts. More complete suppression of the doxastic conclusion would also lead to higher plausibility ratings in the face of sequels inconsistent with them. Almost complete suppression of the weaker doxastic conclusion from ‘look’, which removes the ‘extra’ tension with the sequel, would explain why ‘look’-items are deemed no less plausible than ‘is’-items, while merely partial suppression of such conclusions from ‘appear’ and ‘seem’ would explain why items with these verbs were deemed less plausible than ‘is’-counterparts. While these considerations favour our explanation, we would welcome further research.

4. Philosophical Relevance?

Roberts, Allen and Schmidtke turn from our experimental findings to the question of their philosophical relevance: to what extent do they support our reconstruction of the argument from illusion? According to our analysis, this argument proceeds from a contextually inappropriate stereotypical inference from the appearance verb in the initial premise (e.g., ‘The coin appears elliptical to the viewer’), to a doxastic conclusion (*The viewer thinks the coin is elliptical*). With this input, application of the representativeness heuristic leads to the conclusion that viewer and coin probably do not fall under the category *x is aware of y*, i.e., that the viewer is not aware of the coin. Only this negative conclusion, we argue, renders intuitive the Phenomenal Principle current versions of the argument invoke (see below).

Roberts, Allen and Schmidtke ask an excellent question: Once contextually inappropriate stereotypical inferences like the inference from ‘surgeon’ to male gender in the accident vignette above (Sec. 3) are made explicit, we have no problem to suppress them and disregard their conclusions. The 'surgeon problem' disappears the moment we make explicit the inappropriate inference that generates it. If our analysis of the argument from illusion is correct, this argument should therefore lose its intuitive force, the moment we expose the contextually inappropriate doxastic inferences at its root. But the argument seems to retain its intuitive force. So why is that?

An answer is provided by our Salience Bias Hypothesis, as more fully developed in two subsequent papers which used eye tracking and pupillometry to experimentally examine contextually inappropriate inferences from rare (epistemic) uses of polysemous perception verbs (Fischer & Engelhardt, 2019a; 2019b): We suggest arguments from illusion are driven, more specifically, by contextually inappropriate inferences from rare (phenomenal) uses of polysemous (appearance) verbs. These inferences are supported by the complex, internally structured stereotype or 'situation schema' associated with the verbs' dominant (doxastic) sense. According to our Salience Bias Hypothesis, such inferences occur when the dominant sense is far more salient (i.e., frequent and prototypical) than all other senses or uses and utterances employing the less salient (infrequent) use at issue are interpreted by retaining the dominant situation schema, which is initially activated by default, and attempting to suppress its contextually inappropriate components (Retention/Suppression Strategy; Giora, 2003).

Such attempts remain unsuccessful, and inappropriate inferences influence further cognition, when only some of the frequently co-occurring core schema components are contextually relevant: Since, once activated, they exchange lateral cross-activation (McRae et al., 2005), it is then impossible to suppress some of them completely, while retaining the others that are contextually relevant. Schema components that are only partially suppressed continue to support inferences. Contextually inappropriate stereotypical inferences thus become persistent – and resistant to explicit insight – where unbalanced polysemes are interpreted with a particular interpretation strategy (Retention/Suppression) and applied in contexts where some, but not all of the core components of their complex schemas are relevant. Hence even exposure of the inappropriate inferences does not diminish their intuitive force or that of the argument they drive.

Whereas early 20th century versions of the argument from illusion leap from initial premises ('Viewed sideways, the coin appears elliptical') to the conclusion that the viewer is not aware of the round coin, current versions invoke the Phenomenal Principle ('Whenever something appears F to a subject, the subject is (directly) aware of something that actually is F'). Roberts, Allen and Schmidtke observe that proponents of this principle do not regard it as intuitively plausible in the abstract, but accept it because they find intuitive specific concrete instances ('When a coin looks elliptical to a viewer, the viewer is aware of something that is elliptical'). Proponents have these intuitions, Roberts, Allen and Schmidtke argue, because they attend to particular perceptual experiences and find that 'their phenomenal judgement will support the view that there exists something or other that is, say, brown: that there is a phenomenological similarity to the experience of something that is really brown, and something that merely appears brown in a particular set of circumstances' (as Roberts, Allen and Schmidtke put it). The Phenomenal Principle, they appear to suggest, is to be assessed through phenomenological observation rather than empirical inquiry into automatic inference processes.[3]

[3] In the next paragraph, however, Roberts, Allen and Schmidtke suggest that the 'intuition' or 'view' at issue is supported by inference to the best explanation, a suggestion we discuss in the target paper (Fn.24).

While Roberts, Allen and Schmidtke suggest that 'phenomenological disagreements … cannot be explained in terms of inappropriate stereotypical inferences', we would note that such inferences may colour statements of phenomenological facts, and reasoning from them. The target paper suggests that inappropriate doxastic inferences lead from the initial premise of the argument from illusion to the conclusion that the viewer is not aware of, say, the round coin (above). As the paper then argues (Fischer et al., 2019, Sec.7.3), this negative conclusion leads to a misinterpretation of the powerful intuition that the viewer is aware of an elliptical silvery patch when looking at the round coin sideways – and only this misinterpretation supports inferences in line with the Phenomenal Principle.

The gist of our argument is this. We sometimes use words metaphorically, in referring to things by what they look like: In most non-fiction contexts, 'A ghost opened the door' would be understood not as an endorsement of the supernatural but to mean that someone looking in some ways like a stereotypical ghost (pale as a ghost, disguised as a ghost, etc.) opened the door. We resort to such metaphorical talk, e.g., when we want to avoid the implication that the viewer knew what she was seeing ('She watched the silvery spots in the sky grow bigger and realised too late they were enemy planes'). When appearance-verbs are used to describe cases of non-veridical perception, they prompt attributions of wrong beliefs (*The viewer thinks the coin is elliptical*), which are inconsistent with epistemic implications from 'aware of', and thereby render it natural to say that 'The viewer is aware of an elliptical patch' when looking at the round coin. However, in this metaphorical use, 'elliptical patch' refers to the round coin and the statement does not license the inference that the viewer is aware of something that actually is elliptical.

Inferences in line with the Phenomenal Principle require a literal interpretation of 'elliptical patch'. We submit that proponents of the argument from illusion switch from an uncontroversial metaphorical interpretation to a literal interpretation only because they already presuppose that the viewer is not aware of the round coin. We explain the switch by reference to the partial match heuristic for determining reference which has been invoked to explain semantic illusions (e.g., Park & Reder, 2004): 'Pick the domain element most similar to the stimulus concept, if the similarity exceeds a threshold; otherwise, assume the expression has a reference satisfying the concept, outside the domain of discourse.' As long as the (say) round coin is regarded as part of this domain, 'elliptical patch' is taken to refer to it. But the prior negative conclusion that the viewer is not aware of the round coin excludes this physical object from the domain of discourse (here: things the viewer is aware of). The partial match heuristic thus has proponents of the argument posit a reference that satisfies the description on its default literal reading. We therefore suggested that acceptance of the Phenomenal Principle as intuitive is based on prior acceptance of the negative conclusion that earlier versions of the argument from illusion infer directly from its initial premises.

In contrast with our explanation of how proponents of the argument arrived at the negative conclusion, this account of the Phenomenal Principle remains to be experimentally examined. The empirically informed account may, however, suffice to suggest that it can be philosophically rewarding to complement phenomenological efforts with an empirical investigation of automatic comprehension inferences in phenomenological reports and arguments.[4]

[4] For helpful comments on a previous draft, we thank Joachim Horvath, Hiroshi Ohtani, Justin Sytsma, and Keith Allen.

References

Austin, J.L. (1962). *Sense and Sensibilia*. Oxford: OUP

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

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/

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

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

Fischer, E., & Engelhardt, P.E. (2019a). 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. (2019b). Lingering stereotypes: Salience bias in philosophical arguments. *Mind and Language*. DOI: 10.1111/mila.12249

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

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

Grice, H.P. (1961). The causal theory of perception. *Proceedings of the Aristotelian Society*, 35, 121-152.

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.

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

Machery, E. (2017). *Philosophy within its Proper Bounds*. Oxford: OUP

Park, H., & Reder, L.M. (2004). Moses illusion. In R. Pohl (ed.), *Cognitive Illusions* (pp. 275–291). New York: Psychology Press.

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