The challenge and art of examining conceivability intuitions: Response to Comments[1]

Eugen Fischer and Justin Sytsma

In Fischer and Sytsma (2021) we put forward a bold hypothesis: the zombie argument against materialism is built on *zombie intuitions* – intuitions that are 'killed' (cancelled) by the context provided but kept cognitively alive by linguistic salience bias. We then provided evidence from corpus studies as well as surveys and experiments with typicality, plausibility, and agreement ratings to support this hypothesis. The four commentators have provided helpful and thought-provoking objections, in particular to our main experiment, that point to new hypotheses. Here, we'll respond to the principal points our commentators raise, focusing on the new hypotheses and how they might be tested. We briefly summarise the target article in Sect.1, with a focus on the aspects targeted by commentators. Sect.2 discusses the primary objections Chalmers and Liu raised, namely, to the experimental materials we used, and spells out the competing hypotheses their objections motivate. Sect.3 reports a follow-up study that examined these hypotheses. In Sect.4, we turn to further concerns about the main experiment's materials and procedure, raised by Frankish and Machery. In conversation with these two commentators, the final Sect.5 brings out the need for empirical investigation of laypeople's intuitions about philosophical zombies (and other 'problem intuitions' motivating the 'hard problem of consciousness') and highlights what is new and important about our ambitious 'aetiological strategy' that seeks to develop and assess debunking explanations of intuitions.

1. *Zombie intuitions*

'Zombie intuitions', the target paper, initiates the experimental investigation of conceivability intuitions. This involves methodological challenges. As a pre-study we reported in the paper revealed, few laypeople master the relevant notions of conceivability and contradictoriness sufficiently well to competently answer direct questions about the conceivability of outlined scenarios. The paper therefore used a new paradigm that emulates Chalmers's test for positive conceivability. This *POSCON test* has a thinker try to imagine a situation that verifies a statement S, hypothetically assume that this situation is actual (rather than counterfactual), and intuitively assess whether it follows from that assumption that S is the case. In the experimental implementation, a vignette prompts participants to imagine a pertinent situation and treat it as actual. A subsequent agreement rating task then elicits judgments about whether S is true in (or 'verified by') the imagined situation. These 'verification judgments' provide defeasible evidence of positive conceivability. This evidence can be defeated, e.g., by debunking explanations of the verification intuitions.

The paper uses the new paradigm to examine intuitions about the conceivability of philosophical zombies that are at the root of the 'hard' problem of consciousness. In an extension of the 'negative' research program in experimental philosophy, the paper seeks to develop and support a debunking explanation of the pertinent verification judgments. These judge that certain

[1] We thank David Chalmers, Keith Frankish, Michelle Liu, and Edouard Machery for their helpful and stimulating comments.

imagined beings are physico-behaviourally indistinguishable from average humans (**'P'**) but do not have conscious experience (**'~Q'**; statement of interest S = **P and ~Q**). Linguistic salience bias hits where a familiar word with a clearly dominant sense (like 'zombie') is used in a rare (e.g., philosophical) sense that requires suppression of some, but not all stereotypical features strongly associated with the dominant sense; where this happens, contextually cancelled stereotypical inferences supported only by the dominant sense go through and influence further judgment (Fischer & Engelhardt, 2020).

This bias predicts framing effects in philosophical thought experiments that adapt familiar words to talk about unusual cases – as many philosophical thought experiments do. These include thought experiments about philosophical zombies: After reading a vignette that invites participants to imagine such beings, laypeople will be more likely to attribute to them typical zombie features cancelled by an explicit stipulation of physico-behavioural indistinguishability, when beings are described as 'zombies' rather than 'duplicates'. Pre-studies established that relevant features include *lack of conscious experience*: this is typical of zombies but in tension with physico-behavioural indistinguishability, from which laypeople infer possession of conscious experience. Testing for pertinent framing effects, our main study found (a) that about twice as many participants took the imagined beings to verify both **P** and **~Q** when they were described as 'zombies', rather than 'duplicates'; (b) when these beings were described as 'duplicates', only 15-20% of participants took them to verify both statements. We inferred from (a) that linguistic salience bias influences the prima facie conceivability of philosophical zombies and from (b) that, where suitably neutral formulations are used, philosophical zombies are prima facie conceivable for only a small minority of participants. On this basis, we argued that linguistic salience bias is instrumental for rendering philosophical zombies prima facie conceivable – and that the impression that there is a 'hard' problem of consciousness relies on epistemically deficient intuitions.

We now discuss objections from commentators, which will motivate new hypotheses (Sect.2) and present a follow-up study to assess our previous and the new hypotheses (Sect.3).

2. *'All is dark inside': New hypotheses*

David Chalmers and Michelle Liu accept finding (a) and the conclusion we infer from it – that linguistic salience bias influences the prima facie conceivability of philosophical zombies – but regard them as irrelevant to the assessment of the zombie argument. They regard them as irrelevant, because 'the argument can be formulated without mentioning the word "zombie"' (Liu); indeed, 'people made zombie arguments for years with labels like … "imitation man" instead and they didn't seem notably less effective' (Chalmers). It is, however, an empirical question to what extent the wording of the argument contributes to its 'effectiveness' and, specifically, to the acceptance of its initial conceivability premise. The historical observation that the prominence of the pertinent conceivability arguments seems to have increased considerably upon Chalmers's popularization using the 'zombie' framing does not support the present suggestion. Our study provides quantitative evidence to address the question of 'effectiveness': The finding (a) – which Chalmers and Liu set aside – shows that arguments that speak of 'duplicates', rather than 'zombies', are significantly less effective in persuading at any rate laypeople to accept that philosophical zombies are conceivable.

More important are questions of epistemic warrant. On Chalmers's approach, assumptions of positive conceivability are in need of empirical support: Defeasible evidence is provided by verification judgments; this evidence may be defeated by debunking explanations of these judgments, including explanations that trace them back to cognitive biases. The finding – from (a) – that linguistic salience bias accounts for up to half of all positive verification judgments when arguments speak of 'zombies' then undermines the evidentiary value of these judgments. The finding shows that conceivability assumptions need to be supported by verification judgments that are made about scenarios described in neutral terms (like 'duplicate') – and, preferably, made by judges not previously exposed to similar scenarios speaking of 'zombies'.

Both Chalmers and Liu go on to suggest that our main study failed to establish that, as we inferred from (b), philosophical zombies are not prima facie positively conceivable for a majority of laypeople when the relevant creatures are neutrally described as 'duplicates'.

Chalmers provides two reasons. First, he thinks we asked the wrong questions, since we did not ask the participants in our main study 'whether **P**&**~Q** is conceivable or anything in the vicinity'. This objection strikes us as methodologically problematic: In our pre-study, we did ask participants to rate the conceivability and contradictoriness of philosophical zombies (described as 'duplicates'). Over a third of participants declared themselves agnostic about the scenario's conceivability or its contradictoriness. A quarter declared themselves agnostic about both. A further quarter either agreed or disagreed with both questions (found the scenario both contradictory and conceivable or both non-contradictory and inconceivable). Overall, we found no significant negative correlation between conceivability and contradictoriness judgments. Clearly, laypeople are not sufficiently proficient with the relevant notions to make the elicitation of explicit judgments of conceivability – or contradictoriness (essential to negative conceivability) – a useful format for studying conceivability. Laypersons' answers to direct questions about conceivability bear little evidentiary weight. To obtain empirical evidence that philosophical zombies are prima facie conceivable for laypeople, we need a different approach – like the positive conceivability test set out above.

Chalmers's second reason relates to a dilemma that faces the experimental implementation of this test. The dilemma arises from the prospect of shallow processing (failing to integrate information from different parts of the text; Ferreira et al., 2002):

> **[Dilemma]** 'Including in the vignette, for effectiveness, explicit statements of **P** and **~Q** risks verification judgements based just on recognition that the item appeared in the vignette, without taking other information into account. Conversely, avoiding explicit statements of **P** and **~Q** to prevent [such] shallow processing risks to leave participants without sufficiently clear guidance for the imagination task.' (Fischer & Sytsma, 2021, p.10)

To negotiate this difficulty, our vignette included an explicit statement of **P**, but employed the metaphor 'all is dark inside', to state **~Q**. This is the target of Chalmers's further objection. The 'duplicate' version of our vignette, he maintains, did not press participants hard enough to imagine beings that lack conscious experience: While the 'zombie' version pressed people to do so, through repeated use of this noun, the 'duplicate' version only said that 'When it comes to the duplicate, all is dark inside'. If we had pressed participants harder, he suggests, a higher proportion might have succeeded in imagining beings they were willing to accept as verifying both **P** *and* **~Q**. However, our paper suggested, pressing participants harder might have deprived their judgments of evidentiary value, due to shallow processing.

With a similar ultimate thrust, Liu develops a more specific criticism of our metaphorical statement of **~Q**: To facilitate the intended interpretation of 'when it comes to the duplicate, all is dark inside' (namely, *the duplicate lacks conscious experience*), the vignette needs to provide contextual cues that indicate its target domain (consciousness). In the absence of sufficiently salient cues, participants may interpret the metaphor differently (e.g., our target paper mentioned the alternative interpretation *the duplicate is full of bad thoughts and feelings*); alternatively, participants will then fail to extract any information from the sentence. In either case, the vignette would fail to prompt participants to make an effort to imagine a scenario that verifies both **P** *and* **~Q**. Participants might hence judge that the scenario they imagine does not verify **'~Q'**, simply because they did not even try to imagine a pertinent situation – even if they would have done so successfully if provided with more contextual support for the intended interpretation. Liu helpfully suggests an alternative formulation of the vignette that still employs a metaphor (thus helping to avoid shallow processing) but provides stronger contextual support to facilitate its intended interpretation (thus providing sufficiently clear guidance for the imagination task).

This objection raises the question of how participants interpret the metaphor 'all is dark inside'. The paper's interpretation of results was partially predicated on the hypothesis

H1 Participants interpret 'all is dark inside' as stating that the beings to be imagined are incapable of having conscious experiences, in both the zombie and the duplicate condition.

While the paper did not rely on **H1** in assessing the hypothesis that linguistic salience bias extends to 'zombie', it did assume **H1** in assessing the hypothesis that linguistic salience bias increases the prima facie conceivability of philosophical zombies and in gauging the proportion of participants for whom such zombies are prima facie conceivable. Liu's objection suggests two apparently competing hypotheses:

H2 Participants place a different interpretation on 'all is dark inside', in both conditions.

H3 Participants do not place any determinate interpretation on 'all is dark inside', in either condition.

We would add that the observed stereotypical association between 'zombie' and *lacks conscious experience* (which 'duplicate' arguably lacks) suggests that the use of 'zombie' might render the intended target domain sufficiently salient. This suggestion motivates a fourth hypothesis:

H4 Participants interpret 'all is dark inside' as stating that the beings to be imagined are incapable of having conscious experiences, more frequently in the zombie condition than in the duplicate condition.

This would be consistent with the larger framing effect we observed for consciousness attributions compared to attributions of other typical zombie features (e.g., *has a rotting body and attacks and eats humans*) and atypical zombie features (e.g., *capable of being happy, singing, smelling flowers, and feeling love*): This difference could be due to 'zombie' (but not 'duplicate') facilitating the intended interpretation of the metaphor, thus leading participants to make more effort to imagine a pertinent situation, in the zombie than in the duplicate condition.

3. Experiment

We conducted a follow-up study to provide an initial assessment of these hypotheses.

3.1.Methods

Participants were recruited from the same population (native English speakers raised in North America, over 16 years of age) and with the same strategy (advertising on Google for a free personality test administered after the task) as for the main study in the target paper. Recruitment continued until 100 participants (50 per condition) had passed an attention check.[2]

Materials: Each participant read a version of the text of our previous vignette (in quotes below) that used either 'zombie' or 'duplicate'.

> Here is a brief story:
>
> "In the future scientists are able to exactly scan a person's body, including their brain, at the molecular level. Using this information, they can then create an exact physical duplicate of that person's body and brain, molecule by molecule. The resulting ['zombie'/duplicate] will have a body and brain just like the original person's. The [zombie/duplicate] will also behave just like that person. But, when it comes to the [zombie/duplicate], all is dark inside."
>
> We are interested in what the last sentence means in the context of this story. What do you think the author means when they say, **"When it comes to the [zombie/duplicate], all is dark inside"**? Do you think that each of the following sentences says the same thing, says something completely unrelated, or says the exact opposite thing as **"When it comes to the [zombie/duplicate], all is dark inside"** says in the story?

Participants then used a 7-point Likert scale – anchored at -3 with 'Exact Opposite', at 0 with 'Completely Unrelated', and at 3 with 'Exact Same' – to assess a set of eight sentences. These included sentences probing possible interpretations *ex positivo* and *ex negativo*: They probed the intended interpretation (NCE and IML below), what we regarded as the main competing metaphorical interpretation (MOROSE and HAPPY), and literal interpretations (UNLIT and LIGHTS). Two fillers (STIFFLY and TALKS) were included to allow us to assess whether participants addressed the interpretation question we put to them or rather the different question of whether the presented claim is likely to be true in the situation described (in which case linguistic salience bias predicts higher ratings for STIFFLY and lower ratings for TALKS in the zombie condition compared to the duplicate condition):

NCE:	The [zombie/duplicate] is incapable of having conscious experiences.
IML:	The [zombie/duplicate] has an inner mental life, including feelings and emotions.
MOROSE:	The [zombie/duplicate] is morose and full of bad thoughts.
HAPPY:	The [zombie/duplicate] is happy and cheerful.
UNLIT:	The [zombie/duplicate] is sitting in an unlit room.
LIGHTS:	The [zombie/duplicate] has turned on all the lights in their home.
STIFFLY:	The [zombie/duplicate] moves stiffly.
TALKS:	The [zombie/duplicate] talks.

These sentences were presented on the same page, in random order, along with the attention check already used in the main study ('Please select 1 for this item.') The second page included an open text question:

> You just considered what the sentence "When it comes to the [zombie/duplicate], all is dark inside" means in the short story you read. Using your own words, please paraphrase in the space below what you take to be the intended meaning.

---

[2] Participants were 77.0% women (five non-binary), with an average age of 29.8 years (ranging from 16 to 78 years).

3.2.Results and discussion

We first considered paraphrases provided in response to the open text question. 85 of our 100 participants provided paraphrases. These reflected three recurrent themes, namely, lack of consciousness, morosity, and evil (the last of which had not clearly figured in our items). It also included a ragbag of further metaphorical interpretations. We therefore coded them as 'lacks conscious experience' (39), 'morose' (6), 'evil' (15), 'other metaphorical' (18), 'literal' (2), and 'unclear paraphrase meaning' (9). Four paraphrases received two codes (2 'lacks conscious experience' and 'evil', 2 'morose' and 'evil'). This means that 46% of the provided paraphrases reflected our intended interpretation of 'lacks conscious experience', while the next most frequent interpretation ('evil') was reflected by 18% of paraphrases. However, many 'lacks conscious experience' explanations referred only to paradigmatic conscious experiences (e.g., 'has no feelings' or 'has no emotions and/or thoughts of its own'), rather than explicitly stating that the claim at issue excludes *all* kinds of conscious experiences.

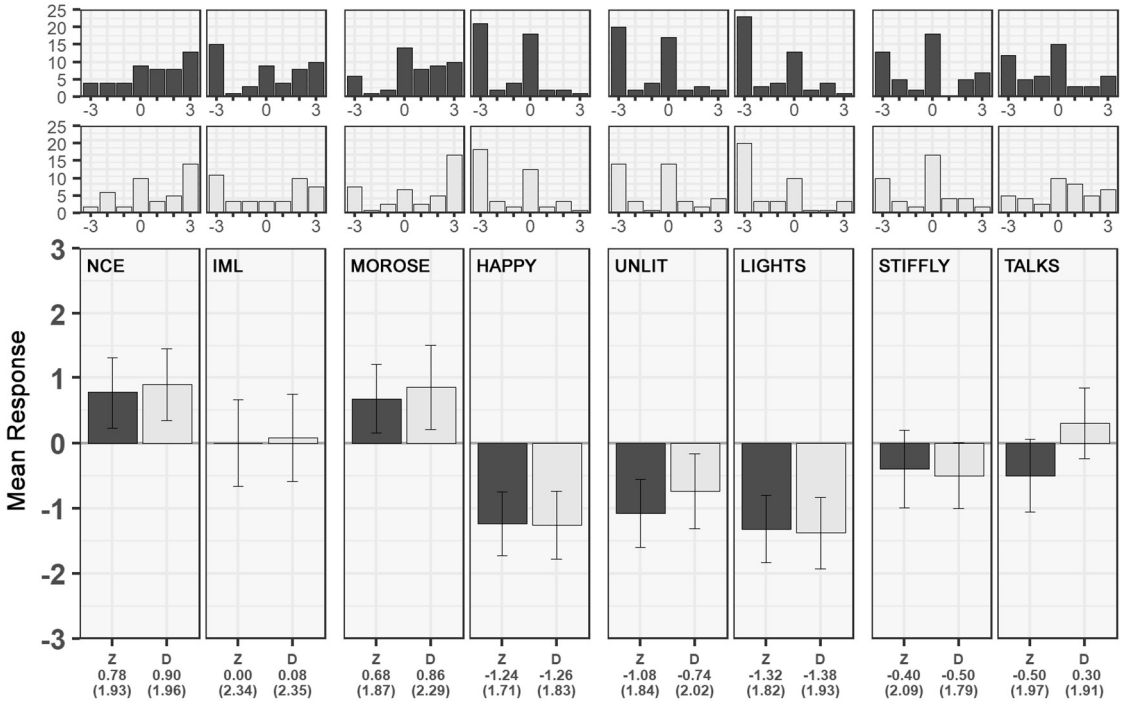

Figure 1: Results with means followed by standard deviations below the bar graphs, for zombie (Z) and duplicate (D) conditions; bar graphs showing 95% confidence intervals. Histograms above each bar graph show the frequency distributions of responses for each condition.

Next we considered the item ratings from the first page. Results are shown in Figure 1. We began by comparing the two conditions. Welch's t-tests for each item applying the Holm–Bonferroni correction for eight comparisons showed no significant differences.[3] We therefore combined conditions except where noted otherwise. The lack of a difference between zombie and

[3] Throughout, all tests reported are two-tailed. Where corrections for multiple comparisons are applied, the uncorrected p-value is shown in square brackets. In the present case, only TALKS was significant without the correction. NCE: $t(97.975)=.31$, $p=1$ [.76], $d=.062$; IML: $t(97.999)=.17$, $p=1$ [.86], $d=.034$; MOROSE: $t(94.115)=.43$, $p=1$ [.67], $d=.086$; HAPPY: $t(97.564)=-.056$, $p=1$ [.96], $d=.011$; UNLIT: $t(97.161)=.88$, $p=1$ [.38], $d=.18$; LIGHTS: $t(97.706)=-.016$, $p=1$ [.87], $d=.032$; STIFFLY: $t(95.695)=-.26$, $p=1$ [.80], $d=.051$; TALKS: $t(97.896)=2.06$, $p=.34$ [.042], $d=.41$.

duplicate conditions for NCE and ILM speaks clearly against **H4**,[4] and we did not consider this hypothesis further.

The eight items tested include two of the three interpretations that recurred in the open text responses, namely NCE and MOROSE. These two were the only items that participants tended to endorse, with mean ratings for these being significantly above the neutral point.[5] These two items were also endorsed (rating of 1, 2, or 3) by the highest proportion of participants across the two conditions (56% for each). The next highest were IML (47% endorsing) and TALKS (36% endorsing) – neither of which occurred in the open text paraphrases. This suggests that participants often endorsed not only items that they took to articulate what the sentence of interest states, but also items that they took to be *entailed* or *implied* by the sentence of interest: MOROSE entails IML, which implies TALKS (verbal statements being a typical manifestation of inner mental life).

To assess this suggestion and gauge the consistency of participants' ratings, we examined key correlations. As predicted by the present suggestion, there were significant positive correlations between ratings for MOROSE and IML (r=0.30, p=.0025) and between ratings for IML and TALKS (r=0.29, p=.0038). Further, over half of the participants who endorsed MOROSE also endorsed IML (32/56=57%) and nearly half of those who endorsed IML endorsed TALKS (23/47=49%). There were also significant negative correlations between ratings for the mutually contradictory items NCE and IML (r=-0.27, p=.0072), MOROSE and HAPPY (r=-0.28, p=.0052), and NCE and MOROSE (r=-0.25, p=.011), suggesting that many participants interpreted these items as being in tension. In line with this, only a minority endorsed both items in any of these pairs, with the percentage being significantly below 50% in each case.[6]

These findings suggest that the bulk of our participants endorsed one of two interpretations frequently perceived as conflicting, namely NCE or MOROSE. The facts that each of these was endorsed by 56% of our participants and that the negative correlation between the ratings for these items remained shy of medium sized effect, however, indicates that some participants endorsed both, and thus failed to arrive at a single interpretation. To assess this, we broke down the patterns of responses for these two items in terms of those endorsing each item.[7] We found that half of the 56 participants who endorsing NCE also endorsed MOROSE.

These findings provide qualified support for each of the apparently competing hypotheses **H1-H3**. Consistent with **H1**, over half (56%) of participants endorsed the intended interpretation NCE, and just under half (46%) of freely produced paraphrases reflected this interpretation. However, only just over a quarter of participants (28%) endorsed this interpretation without simultaneously endorsing the competing interpretation MOROSE, and

---

[4] In line with this, the proportion of paraphrases corresponding with NCE was virtually identical for each condition (44% of produced paraphrases in the zombie condition, 48% in the duplicate condition).

[5] We conducted a series of one-sample t-tests to compare ratings for each item to the midpoint, again applying the Holm–Bonferroni correction for eight comparisons (uncorrected p-value shown in square brackets). NCE: t(99)=4.34, p<.001 [<.001], d=.43; IML: t(99)=.17, p=1 [.86], d=.017; MOROSE: t(99)=3.70, p=.0014 [<.001], d=.37; HAPPY: t(99)=-7.10, p<.001 [<.001], d=.71; UNLIT: t(99)=-4.72, p<.001 [<.001], d=.47; LIGHTS: t(99)=-7.23, p<.001 [<.001], d=.72; STIFFLY: t(99)=-2.33, p=.066 [.022], d=.23; TALKS: t(99)=-.51, p=1 [.61], d=.051.

[6] NCE and IML: 22%, χ^2=30.25, p<.001; MOROSE and HAPPY: 6%, χ^2=75.69, p<.001; NCE and MOROSE: 28%, χ^2=18.49, p<.001

[7] 28% of participants endorsed both NCE and MOROSE, 28% endorsed NCE but not MOROSE, 28% endorsed MOROSE but not NCE, and 16% endorsed neither.

thus won through to placing *only* our intended interpretation on the phrase of interest. Consistent with **H2**, over half (56%) of participants endorsed the alternative interpretation MOROSE. Even though only 7% of produced paraphrases reflected this interpretation, just over a quarter (28%) of participants endorsed this interpretation without simultaneously endorsing the intended interpretation NCE. Responses to the paraphrasing task revealed a further interpretation ('evil', 18% of produced paraphrases) that is compatible with, and can complement, both the intended and the alternative interpretation. Consistent with **H3**, over a quarter of participants (28%) endorsed two interpretations (NCE and MOROSE) that were frequently perceived as in tension, while 11% of the paraphrases provided were utterly garbled. This suggests that a not negligible number of participants failed to win through to a determinate interpretation of the sentence of interest.

So where does this leave the findings from Fischer and Sytsma (2021)? They continue to support the crucial hypothesis that linguistic salience bias affects judgments about philosophical zombies: We observed framing effects for attributions of typical and atypical zombie properties (T1-T3 and A1-A3 in the main study) that are (T1-T3) cancelled by vs (A1-A3) consistent with our vignette's literal statement of physico-behavioural indistinguishability (**P**). These effects provide evidence of the bias, namely, of contextually cancelled stereotypical inferences influencing further cognition – irrespective of how participants interpret the vignette's metaphorical statement of lack of conscious experience (**~Q**). This finding, that linguistic salience bias affects judgments about philosophical zombies, provides support also for the more specific hypothesis that this bias affects the prima facie positive conceivability of philosophical zombies.

To be able to quantify this effect, however, and to determine the proportion of participants for whom philosophical zombies are prima facie positively conceivable, we need to restrict attention to those participants who make an effort to imagine a situation that verifies both **P** *and* **~Q**. Arguably, only participants who place the intended interpretation (NCE) on 'all is dark inside' will do so. While (NCE) was the most popular interpretation across paraphrase and agreement rating tasks, present findings thus suggest that at most just over half the participants made the relevant effort. To quantify the effect and proportion of interest, we therefore need to restrict our attention to the verification judgments of those participants who interpret the metaphor as intended – and, ultimately, to the judgments of those who place on the metaphor the intended interpretation and no interpretation in tension with it. This requires the prior identification of the relevant participants, e.g., through the interpretation tasks used in the present follow-up study.

4. Revisiting materials and procedure

Our target paper brought out the need to move from direct questions about conceivability to less direct approaches like the POSCON test, to empirically document conceivability intuitions. The paper pointed out a dilemma facing empirical implementations of the POSCON test, explored one way of addressing it, and invited exploration of further routes that might be taken. Our commentators have helpfully identified different difficulties that need to be borne in mind. As discussed in Sect.2, Chalmers and Liu question our use of the metaphor 'all is dark inside' to state **~Q** in the vignette inviting participants to imagine beings that qualify as philosophical zombies.

Keith Frankish raises a related concern about the wording of the vignette:

> The key point is that in imagining a philosophical zombie it is not enough to imagine a physical duplicate. We must imagine a *bare* physical duplicate—a duplicate in all physical respects *and no others*. We must not imagine a *rich* duplicate, which has properties that normally accompany these physical properties in virtue of contingent natural laws, causal or otherwise. (p.2)

Our vignette had introduced the entity to be imagined as 'exact physical duplicate' of a 'person's body and brain', but did not explicitly state that duplication did *not* extend further, beyond the physical. It is therefore fair to suggest, with Frankish (p.2), that participants might have imagined a 'rich duplicate'. The study of course addresses the question, to what extent people are able to imagine bare duplicates, rather than just rich ones. The concern, however, is once again whether our vignette provided clear enough guidance to participants about what they should try to imagine: Agreement with consciousness attributions might considerably decline, if vignettes make explicit that in the scenario to be imagined the duplication does *not* extend beyond the physical.[8]

Table 1. Mean ratings and comparisons against neutral mid-point 4 for the restricted sample (N=33) from Sytsma and Snater (in press) using one-sample Student's t-tests [p-value after Holm-Bonferroni correction for 25 comparisons]. Key items highlighted.

Item	Mean	Student's t-tests (mu=4) **[adjusted]**
[1] The duplicate would have free will.	**5.45**	$t(32)=4.63$, $p<.001$ **[<.001]**, $d=0.81$
[2] The duplicate would feel pain when she is injured.	**5.00**	$t(32)=2.85$, $p=.0076$ **[.038]**, $d=0.50$
[3] The duplicate would see colors.	**6.12**	$t(32)=8.04$, $p<.001$ **[<.001]**, $d=1.40$
[4] The duplicate would experience sights and sounds.	**5.79**	$t(32)=6.44$, $p<.001$ **[<.001]**, $d=1.12$
[5] The duplicate would make choices.	**5.79**	$t(32)=6.69$, $p<.001$ **[<.001]**, $d=1.16$
[6] The duplicate would understand English.	**5.58**	$t(32)=5.28$, $p<.001$ **[<.001]**, $d=0.92$
[7] The duplicate would give meaningful replies to questions in English.	**4.94**	$t(32)=2.55$, $p=.016$ **[.064]**, $d=0.44$
[8] The duplicate would display creativity.	**5.42**	$t(32)=4.49$, $p<.001$ **[.0013]**, $d=0.78$
[9] The duplicate would think.	**5.79**	$t(32)=6.44$, $p<.001$ **[<.001]**, $d=1.12$
[10] The duplicate would solve problems.	**5.52**	$t(32)=5.13$, $p<.001$ **[<.001]**, $d=0.89$
[11] The duplicate would be capable of morality.	**4.82**	$t(32)=2.48$, $p=.019$ **[.064]**, $d=0.43$
[12] The duplicate would have dreams when she sleeps.	**4.64**	$t(32)=1.67$, $p=.11$ **[.21]**, $d=0.29$
[13] The duplicate would have emotions.	**5.30**	$t(32)=4.06$, $p<.001$ **[.0029]**, $d=0.71$
[14] The duplicate would have moods.	**5.36**	$t(32)=4.44$, $p<.001$ **[.0014]**, $d=0.77$
[15] The duplicate would have self–consciousness.	**5.00**	$t(32)=3.19$, $p=.0032$ **[.022]**, $d=0.55$
[16] The duplicate would have a personality.	**5.24**	$t(32)=4.08$, $p<.001$ **[.0029]**, $d=0.71$
[17] The duplicate would be intelligent.	**5.33**	$t(32)=4.80$, $p<.001$ **[<.001]**, $d=0.84$
[18] The duplicate would be alive.	**5.03**	$t(32)=3.12$, $p=.0038$ **[.023]**, $d=0.54$
[19] The duplicate would be conscious.	**5.09**	$t(32)=3.53$, $p=.0013$ **[.010]**, $d=0.61$
[20] The duplicate would be aware of herself.	**5.33**	$t(32)=4.35$, $p<.001$ **[.0017]**, $d=0.76$
[21] The duplicate would be aware of things around her.	**5.73**	$t(32)=6.68$, $p<.001$ **[<.001]**, $d=1.16$
[22] The duplicate would be responsible for her actions.	**5.67**	$t(32)=5.28$, $p<.001$ **[<.001]**, $d=0.92$
[23] The duplicate would deserve human rights.	**5.39**	$t(32)=4.10$, $p<.001$ **[.0029]**, $d=0.71$
[24] The duplicate would have goals and ambitions.	**5.24**	$t(32)=4.25$, $p<.001$ **[.0020]**, $d=0.74$
[25] The duplicate would have memories of past events.	**4.39**	$t(32)=0.98$, $p=.33$ **[.33]**, $d=0.17$

This concern has been to some extent addressed by a recent study on intuitions about 'bare physical duplicates' (to use Frankish's term): Sytsma and Snater (in press) used a vignette that, like the target paper's, described future scientists duplicating people's bodies and brains, molecule for molecule. The vignette then explicitly added:

[8] Frankish's dense discussion also explicitly or implicitly raises two further concerns, which we discuss below (one of them in response to Edouard Machery who raises it most clearly).

> Since the scientists are only able to scan the person's physical structure, if there was any non-physical aspect to the person—such as a non-physical soul or mind—the scientists would not be able to duplicate that aspect of the person.

Participants used a 7-point Likert scale to rate their agreement with attributions of several mental features to these bare duplicates. While Sytsma and Snater employed a large, global sample (N=886), for purposes of comparison we'll restrict attention to participants that match the sample from the target paper (native English speakers, raised in North America, non-philosophers; N=33), although findings are comparable for the full sample. Basic results with these restrictions are reported in Table 1. Attributions of the features philosophers typically treat as hallmarks of phenomenal consciousness, including **[2]**, **[3]**, **[4]**, and **[13]**, attracted mean agreement ratings significantly above mid-point, as did the key item **[19]**, which ascribed consciousness. Indeed, the ratings for this item (M=5.09) were right in line with the mean ratings for consciousness items (A=5.02, B=5.01, C=5.05) in the duplicate condition in Fischer and Sytsma (2021). The explicit requirement that duplication does not extend beyond the physical does not seem to notably affect responses.

Frankish simultaneously raises a second concern about the use of the word 'duplicate' as a contrast term for 'zombie', which unlike the previous objections intends to challenge whether the target paper provides evidence of linguistic salience bias in the first place. '[I]t may be', he suggests (p.2), 'that the term "duplicate" also has a biasing effect… [I]t may be that the word 'duplicate' tends to evoke a rich reading.' While he offers this suggestion in articulating the previous concern about bare vs rich duplicates, the distinct rationale is worth making explicit: As lexical databases and dictionaries explain, a 'duplicate' is 'a copy that corresponds to an original exactly' (*WordNet*), 'one of two or more specimens of anything exactly or virtually alike' (*OED*), 'either of two things exactly alike and usually produced at the same time or by the same process' (*Merriam-Webster*). The term stereotypically implies the copy is exact in *all* respects – or at any rate all respects covered by the production process. If so, 'duplicate' need not be a neutral contrast term: Even when first introduced as part of the phrase 'physical duplicate', the word might bias participants towards agreeing with attributions of consciousness, and the framing effects we observed for such attributions might be due to lexical features of either 'zombie' *or* 'duplicate'. While it is hard to see how stereotypical inferences from 'duplicate' (rather than from 'zombie') could contribute much to framing effects observed for attributions of typical and atypical zombie properties like *has a rotting body and attacks and eats humans* (T1-T3 and A1-A3), such inferences might partially account for the larger effects we observed for consciousness attributions, which would then be medium-sized only because of the biasing contrast term.

Initial evidence against this concern is provided by the above findings from Sytsma and Snater (in press): They included in the vignette the quoted passage that explicitly cancels any stereotypical inferences from 'duplicate of person' to attributions of non-physical mental features. This elicited the same mean agreement ratings as the duplicate version of the target paper's vignette that did not include this cancellation (and just introduced the duplicate in question as 'exact physical duplicate'). Of course, linguistic salience bias leads to acceptance of contextually cancelled stereotypical inferences. However, all extant studies of the bias observed some influence of contextual cancellation, even though defeated inferences continued to influence judgments (see, e.g., review in Fischer & Engelhardt, 2020). Present findings therefore speak

against the suggestion that ratings for consciousness attributions in the duplicate condition might be affected by linguistic salience bias.

The suggestion can be further assessed by corpus analyses examining the use of 'duplicate'. If (as we suspect) the term is predominantly applied to physical objects that are produced alike, it will only support stereotypical inferences to likeness in relevant physical respects, not to all respects: e.g., from 'duplicate of John's silver coin' to *is silver* but not *is owned by John*. Finally, Frankish's second worry could be addressed by follow-up studies that seek to replicate our findings with the use of alternative contrast terms like 'copy' and 'clone' (which carry more unwanted baggage, but may be the next-best alternatives to 'duplicate'). This will help identify the best, or most neutral, contrast term to use to study framing effects in thought experiments about philosophical zombies.

Finally, Frankish's concern that participants' might imagine 'rich duplicates' with 'properties that normally accompany these physical properties in virtue of contingent natural laws, causal or otherwise' hints at a sound 'technical point' clearly raised as such by Edouard Machery. Machery notes that our vignette places the relevant scenario in the future of the actual world. This means that participants will assess whether philosophical zombies are conceivable under nomological constraints, not just under the logical/conceptual constraints relevant for the kind of epistemic conceivability that is to warrant conclusions about logical or metaphysical possibility. Since all parties to the philosophical debate agree that philosophical zombies are not nomologically possible, the low proportion of participants who agreed that the scenario verifies both **P** and **~Q** might be due to this failure to cancel nomological constraints which might underwrite inferences from **P** (human body and behaviour) to **Q** (conscious experience).

In response, we observe that our lay participants were scientifically innocent.[9] In consequence, their understanding of the workings of the human mind and body was as shallow as the shallow understanding of complex causal systems documented for lay adults by Rozenblit and Keil (2002). This means that, for reasons of ignorance, our participants' judgments are unlikely to take into account nomological constraints – even if these are not cancelled by vignette content or instructions. We therefore assume that the inaccuracy in our implementation of the POSCON test (which we opted for to increase accessibility) does not affect responses. This assumption can be examined by follow-up studies that implement the test more *à la lettre*, e.g., by presenting lay participants with 'a story that tries to describe things that, for all you know, might happen' and asking them to 'imagine that this story was true, and these things actually happened', before asking participants to judge whether statements of interest are true of or in the situation they imagine.

5. Relevance and novelty

In closing, we discuss the philosophical relevance and novelty of the target research, in conversation with Machery and Frankish. We bring out the need for empirical investigation of conceivability assessments, engage with the 'expertise defence' against evidential experimental

[9] In our less restrictive sample (N=247), roughly 75% of participants indicated no training in either the natural sciences or the brain sciences and psychology: Less than 10% had or were majoring in a natural science subject or taking some graduate classes, and only around 6% had or were majoring in psychology or brain sciences, or taking some graduate classes in these subjects.

philosophy, as it could be levelled against the target paper, comment on the advantages of its ambitious aetiological strategy for evidential experimental philosophy, and consider a suggestion about the purpose of the zombie argument that would make aetiological insights into the workings of linguistic salience bias relevant also for a rather different philosophical project.

5.1.Empirical investigation and the expertise defence

On Chalmers's helpful explication of notions of epistemic conceivability, the claim that some entity X is positively conceivable relies on a hypothesis that is clearly empirical in nature. Indeed, we find this explication so helpful precisely because it renders otherwise typically unsubstantiated conceivability claims empirically tractable: Spelled out, the claim of positive conceivability is that evidence is (i) provided by prima facie conceivability and (ii) not defeated by debunking explanations or other factors. The first part, (i), relies on the hypothesis that X *is* prima facie conceivable, namely, for some population P: Members of P will pass the POSCON test for X. This hypothesis needs experimental support. Moreover, wherever (ii) potential defeaters include proposed debunking explanations, these will need to be assessed empirically.

Experimental work of the kind we conducted and proposed is hence required to establish positive conceivability, within Chalmers's two-stage dialectic. It is required to assess the first premise of conceivability arguments – and to get, e.g., the zombie argument off the ground, in the first place. Unlike some of our commentators, we accordingly do not take the upshot of their objections to be that 'for now, the conceivability of philosophical zombies stands undefeated' (as Liu puts it). Rather, we take the joint upshot of our work and their objections to be that evidential support for this conceivability remains to be provided and is then likely to be defeated – and that direct evidence both for prima facie conceivability and for its defeat are even harder to come by than one might think.

Once we are clear on the two-fold need for empirical investigation, the obvious follow-up question asks, which population needs to be investigated for this purpose, at each of the two stages? Our target paper tested laypeople. Advocates of the zombie argument might instead contend that it is really philosophers that matter here. Machery raises this worry on behalf of 'zombie lovers', suggesting that they are 'unlikely to feel threatened by findings about lay people without training in philosophy' (p.2). Specifically, they might appeal to the 'expertise defence' (of which Machery is a staunch critic). This defence of armchair philosophy against evidential experimental philosophy attacks the latter's typical reliance on lay participants: According to the empirically tractable versions of this defence, philosophers benefit from more experience with thought experiments or better conceptual skills or better conceptual resources than laypeople. As a result, different versions of the defence allege, philosophers' judgments about thought experiments are more accurate, or more stable, or less susceptible to cognitive biases like the linguistic salience bias. We now consider various ways in which advocates of the zombie argument could deploy the expertise defence against the experimental work we reported and proposed.

Proponents of the zombie argument typically do not wish to deny the relevance of lay judgments *tout court*: They standardly assume that the zombie argument (and the hard problem more generally) taps into aspects of the folk conception of consciousness. Indeed, this is what is thought to give the problem much of its sting. But advocates of the argument could make a *selective* appeal to the expertise defence. As Machery suggests, they might contend that

philosophers benefit from better methodological or conceptual skills than laypeople, and infer that philosophers will be less susceptible to the linguistic salience bias we've demonstrated for 'zombie'. In other words, advocates of the zombie argument might grant that laypeople are a relevant population for assessing whether philosophical zombies are prima facie conceivable, as in (i) above, but deny that evidence of a defeater among laypeople, as in (ii) above, is relevant to assessing the *philosophical* argument.

In line with findings for other biases (reviews: Machery, 2017; Sytsma & Livengood, 2015), however, there is evidence that philosophers are also susceptible to linguistic salience bias. A recent study by Fischer, Engelhardt, & Herbelot (RR) on contextually cancelled inferences from perception verbs included a large sample of professional academic philosophers, mainly from 14 leading UK philosophy departments. They found that expert philosophers are no less susceptible to linguistic salience bias than psychology undergraduates. As such, absent evidence to the contrary, we shouldn't assume that philosophers are immune from the linguistic salience bias found for 'zombie'.

Conceivability intuitions that are invoked to support modal intuitions about the metaphysical possibility of philosophical zombies belong to several 'problem intuitions' that motivate the hard problem of consciousness. Our study complements previous studies that found a low prevalence of problem intuitions among laypeople (Diaz, 2021; Gottlob & Lombrozo, 2018; Peressini, 2014). In response, proponents of the hard problem could rethink their reliance on lay judgments: They could argue that while human beings are generally familiar with the phenomenon of phenomenal consciousness through their own introspection,[10] only philosophical theorizing conceptualizes the phenomenon in a way that brings out the aspects that place it in principle beyond the reach of science. According to this argument, philosophers benefit from better conceptual resources than laypeople, which improve their ability to conceive of the relevant scenarios. Therefore, proponents of the hard problem could conclude, in line with another version of the expertise defence, that only the problem intuitions of expert philosophers matter. Perhaps such a difference in conceptual resources partially accounts for the discrepancy between the 40% of our participants who find philosophical zombies prima facie conceivable in the zombie condition, and the 60% of philosophers who pronounced such zombies (described as 'zombies') conceivable in Bourget and Chalmers's (2014) survey.[11]

This line of argument, however, raises the question of whether proponents of the hard problem should prefer such conceptual resources from philosophy to what conceptions of conscious experience laypeople might have formed on the basis of their intimate familiarity with the phenomenon. If the arguments motivating the hard problem crucially rely on technical philosophical notions, the arguments would bring out that these philosophical notions – and only these notions – place conscious experience beyond the reach of science. While there are perfectly straightforward psychological motivations for adopting notions that posit entities or properties beyond the reach of science – think of the comfort one may derive from embracing certain

[10] This too is an empirical claim. As Machery notes, there is reason to doubt that laypeople generally conceptualize conscious experience as philosophers do (e.g., Sytsma and Machery 2010, Sytsma and Snater in press).

[11] A variety of aspects of social learning likely further contribute to this outcome.

theological notions – legitimate philosophical motivations for dismissing lay notions of conscious experience in favour of the philosophical conceptions at issue would be harder to fathom.

To sum up, insofar as the hard problem is motivated via appeals to common thinking among laypeople, experimental work on lay participants is relevant for assessing both components of the positive conceivability claim from which zombie arguments proceed: (i) to assess whether philosophical zombies are prima facie conceivable for the population that matters, and (ii) to assess debunking explanations of the relevant verification judgments. It might be tempting at this point to abandon appeals to common thinking. But there is a reason that ordinary concepts and intuitions have been appealed to in these debates. Absent this, we seem to be left with a theoretical notion of consciousness that stipulates something beyond the reach of science, without much reason to take such a notion seriously. A compelling and worthwhile problem would seem to arise here only to the extent that lay conceptions of consciousness place its explanation beyond the reach of science – and the lay conceptions are not themselves readily debunked.

5.2. Aetiology

Our efforts in the target paper are part of a distinctive new extension of experimental philosophy: Machery notes that the target paper may 'push the negative program of experimental philosophy into new territories', namely, by considering the aetiology of the judgments of interest (p.1). But, he wonders (p.2), how much is gained by this novel focus, philosophically? This is an important question. The brief answer pertaining to the specific aetiology of linguistic salience bias is this: Linguistic salience bias leads to framing effects. The documentation of framing effects merely tells us that we cannot trust just any intuition about the given topic. By contrast, explanations of framing effects, e.g., in terms of linguistic salience bias, can adjudicate between frames and tell us which intuitions, elicited by which frame, to discard, and which to potentially take into account. The linguistic salience bias account of intuitions about philosophical zombies tells us that the 'zombie' frame (which Chalmers's response continues to recommend for use) is illicit.

This 'aetiological strategy' can be combined with other critical strategies. The target paper combined the linguistic salience bias account with considerations from the epistemology of peer disagreement, to argue from the finding that only a small minority of participants found philosophical zombies prima facie conceivable in the – arguably legit – duplicate condition, to the conclusion that the degree of peer disapproval undermines the prima facie evidence for the positive conceivability of philosophical zombies that was provided by the minority judgments. Our use of this argument motivated Machery's question. So what is gained by engaging with psycholinguistics, rather than just arguing from peer disapproval? At any rate in the present case, insight into the aetiology paved the way for arguments from peer disagreement: In the zombie condition, 40% of participants appeared to pass the POSCON test. That is arguably too high a proportion to run our arguments from the epistemology of peer disagreement. The linguistic salience bias account then allowed us to discard intuitions elicited in that condition as untrustworthy. Only this move reduced the proportion of apparent conceivers to that small minority that allowed us to run the peer disagreement arguments. Besides, in this case, as more often, only some understanding of the aetiology of divergent judgments renders plausible the contention that the dissenters are epistemic peers about the matter at hand, given the usual –

limited and generic – demographic data about participants (about their highest degree, etc.). Aetiological considerations play a key role in our argument, also where they facilitate further argumentative strategies that may complement debunking-explanation strategies.

While the adoption of the aetiological strategy, and its combination with other critical strategies, represents an innovative extension of negative experimental philosophy, it retains a conventional meta-philosophical perspective that regards philosophical thought experiments as instruments – if potentially defective instruments – for garnering evidence for or against philosophical theories, or revealing truths of philosophical interest. Frankish concludes by suggesting that this might not be the best way to think of the *true* purpose of the zombie thought experiment. He suggests that while the thought experiment purports to be 'revealing the implications of our intuitive grasp of the nature of consciousness', it is actually 'encouraging us to form and endorse a particular *theory* of consciousness—a theory that treats consciousness as a psychic essence distinct from all functional processes' (p.3). Taking this to be the purpose, Frankish suggests that the term 'zombie' is in fact well-suited to the task, in part *because* of linguistic salience bias! When thought experiments are seen as a tool for revealing the truth, linguistic salience bias that asserts itself in the interpretation of vignettes is a *bug*. By contrast, when these thought experiments are seen as a rhetorical tool to elicit agreement, any contextually cancelled default inferences that help with this are better thought of as a *feature*. Accepting this shift – which, of course, we suspect 'zombie lovers' will want to deny – the target paper could then be employed for the rather different project of detailing how a bit of philosophical propaganda gets its rhetorical appeal.

6. Conclusion

The target paper 'Zombie intuitions' commenced the empirical investigation of conceivability intuitions and extended the negative research program in experimental philosophy by complementing the documentation of framing effects by their explanation. The study of conceivability intuitions about philosophical zombies poses several challenges in connection with the formulation of vignette and instructions. In response to helpful comments, the present paper presented two new studies and outlined plans for others that jointly address these challenges. We finally argued that the empirical study of conceivability intuitions is indispensable for motivating and assessing conceivability arguments.

References

Bourget, D., & Chalmers, D. J. (2014). What do philosophers believe? *Philosophical Studies, 170*, 465–500.

Diaz, R. (2021). Do people think consciousness poses a hard problem? Empirical evidence on the meta-problem of consciousness. *Journal of Consciousness Studies*, 28 (3–4), 55–75.

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

Fischer, E., & Engelhardt, P. E. (2020). Lingering stereotypes: Salience bias in philosophical argument. *Mind & Language, 35*, 415–439.

Fischer, E., Engelhardt, P. E., & Herbelot, A. (2021). *The expertise objection: A psycholinguistic perspective*. University of East Anglia (ms).

Fischer, E., & Sytsma, J. (2021). Zombie intuitions. *Cognition*, 215.

Gottlieb, S., & Lombrozo, T. (2018). Can science explain the human mind? Intuitive judgments about the limits of science. *Psychological Science, 29*, 121–130.

Machery, E. (2017). *Philosophy in its Proper Bounds*. OUP

Peressini, A. (2014). Blurring two conceptions of subjective experience: Folk versus philosophical phenomenality. *Philosophical Psychology*, 27(6), 862–889.

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: an illusion of explanatory depth. *Cognitive Science*, 26, 521–562.

Sytsma, J., & Livengood, J. (2015). *The Theory and Practice of Experimental Philosophy*. Broadview Press.

Sytsma, J., & Machery, E. (2010). Two Conceptions of Subjective Experience. *Philosophical Studies*, 151(2): 299–327.

Sytsma, J., & Snater, M. (in press). Consciousness, Phenomenal Consciousness, and Free Will. In P. Henne & S. Murray (eds.), *Advances in Experimental Philosophy of Action*, Bloomsbury. Preprint: http://philsci-archive.pitt.edu/19556/

University of East Anglia
e.fischer@uea.ac.uk

Victoria University of Wellington
jmstsma@gmail.com