Zombie Intuitions

Final draft version of paper forthcoming in *Cognition* (accepted June 7th, 2022)

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

Justin Sytsma
Victoria University of Wellington

In philosophical thought experiments, as in ordinary discourse, our understanding of verbal case descriptions is enriched by automatic comprehension inferences. Such inferences have us routinely infer what else is also true of the cases described. We consider how such routine inferences from polysemous words can generate zombie intuitions: intuitions that are 'killed' (defeated) by contextual information but kept cognitively alive by the psycholinguistic phenomenon of linguistic salience bias. Extending 'evidentiary' experimental philosophy, this paper examines whether the 'zombie argument' against materialism is built on zombie intuitions. We examine the hypothesis that contextually defeated stereotypical inferences from the noun 'zombie' influence intuitions about 'philosophical zombies'. We document framing effects ('zombie' vs 'duplicate') predicted by the hypothesis. Findings undermine intuitions about the conceivability of 'philosophical zombies' and address the philosophical 'hard problem of consciousness'. Findings support a deflationary response: The impression that *principled* obstacles prevent scientific explanation of how physical processes give rise to conscious experience is generated by philosophical arguments that rely on epistemically deficient intuitions.

Experimental philosophy, philosophical intuitions, comprehension inferences, zombie argument, 'hard' problem of consciousness, meta-problem of consciousness.

1. Introduction

Words matter. Sometimes, they matter in ways they should not. We will investigate how logically equivalent formulations in philosophical thought experiments promote different inferences that lead to different judgments about hypothetical cases. 'Evidential' experimental philosophy uses the exposure of such 'framing effects' to argue against reliance on thought experiments, in philosophy and beyond (for reviews, see Machery, 2017, pp.77-89; Mallon, 2016). We extend this research program to challenge an influential thought experiment that suggests there is a 'hard problem of consciousness' – a principled explanatory gap between physical facts and conscious experience that prevents scientific explanations of why and how physical processes give rise to conscious experience (Chalmers, 1996; Levine, 1983). This problem has dominated philosophical discussion of consciousness for more than twenty years and attracted attention from across cognitive science and from the general public. The problem arises from supposedly widely shared intuitive judgments (Chalmers, 2018). The most fundamental of these judgments are conceivability judgments, in particular judgments concerning the conceivability of 'philosophical zombies' (beings that are physico-behaviourally identical to us, but lack conscious experience). We ask whether, and why, these conceivability intuitions are vulnerable to framing effects.

We develop a psycholinguistic account that identifies a previously unrecognised source of framing effects: the linguistic salience bias (Fischer & Engelhardt, 2017; 2019; 2020; Fischer, Engelhardt, & Sytsma, 2020) that affects inferences from words with distinct, but related senses (polysemes) – which account for about 40% of English words (Byrd et al. 1987). Linguistic salience bias generates *'zombie intuitions'*: intuitions that are 'killed' (defeated) by contextual information but kept cognitively alive by the bias. We conducted three corpus studies, four surveys, and an experiment to develop and assess the hypothesis that conceivability intuitions about philosophical zombies are zombie intuitions – namely, vulnerable to framing effects that are due to linguistic salience bias. Whereas the mere exposure

of framing effects only supports the conclusion that the intuitive judgments at issue are unreliable, explanations invoking salience bias help adjudicate between conflicting judgments.

Our study extends 'evidential' experimental philosophy by identifying a previously unrecognised source of framing effects, extending their investigation to conceivability judgments (that are challenging to study experimentally), and demonstrating the usefulness of the approach by debunking conceivability intuitions at the root of a prominent philosophical problem. We now review the 'hard problem of consciousness' (Sect. 1.1), the influential 'zombie argument' at the root of the problem (Sect. 1.2), and the linguistic salience bias we suggest is implicated in the intuitions on which this argument builds (Sect. 1.3). The findings of our main study (Sect. 2) will challenge the core argument for the existence of the 'hard problem' – and illustrate consequences for the methodology of philosophical thought experiments, quite generally (Sect. 3).

1.1. The 'hard problem of consciousness'

The 'hard problem of consciousness' is the problem of explaining 'why and how … physical processes in the brain give rise to conscious experience' (Chalmers, 2018, p.6; *cf.* Chalmers, 1996; Nagel, 1974; Levine, 1983). We can distinguish the *scientific* 'hard problem' of developing pertinent explanations (review: Wu, 2018) and the *philosophical* 'hard problem' of explaining how such explanations are *as much as possible* in the light of philosophical 'problem intuitions' which suggest that any scientific explanation is bound to fall short of what is required (Block, 1995; Sytsma 2009).

The most fundamental of these 'problem intuitions' are modal intuitions. Modal intuitions assert it is logically possible that physical events which actually co-occur with conscious experiences fail to be accompanied by conscious experiences (in philosophical zombies) or co-occur with different experiences (in inverted-spectrum cases) (Chalmers, 2018, p.12). In the standard dialectic (e.g., Chalmers, 1996, pp.93-108), modal judgments support (i) explanatory and (ii) metaphysical 'problem intuitions': (i) If it is possible that certain physical events occur in the absence of conscious experience (in philosophical zombies), then their explanation cannot explain why these events actually go with conscious experience; and (ii) if facts about conscious experience are not logically determined by facts about the physical world, their reductive explanation is impossible, and we cannot identify conscious with physical phenomena (at any rate, not with the standard strategy). Modal intuitions are therefore the most fundamental of the 'problem intuitions' that generate the philosophical 'hard problem'. While labelled 'intuitions', modal judgments are, in turn, defended by so-called 'conceivability arguments' that proceed from intuitions about the conceivability of philosophical zombies (or inverted spectra). These conceivability intuitions constitute the ultimate intuitive foundation of the philosophical 'hard problem'.

The hard problem can be rendered empirically more tractable by considering why thinkers think there is a hard problem (or why they think it is so hard), if in fact they do. This question poses the 'meta-problem of consciousness' (for reviews, see Chalmers, 2020a; 2020b; Sekowski & Rorot, 2021). It calls for an explanation of the relevant problem intuitions, in 'topic-neutral' terms that do not invoke phenomenal consciousness (Chalmers, 2018, p.16). Debunking explanations of problem intuitions, which reveal we have no warrant to accept them, will help 'dissolve' the philosophical hard problem, namely, help show that we have no right to believe there are *principled* obstacles to scientific explanations of why and how physical processes give rise to conscious experience. We target conceivability intuitions about philosophical zombies (Sect. 1.2) and develop a 'topic-neutral' psycholinguistic explanation to debunk these intuitions (Sects. 1.3-1.4).

1.2. The zombie argument

The zombie argument invokes conceivability intuitions to argue for the logical possibility of *'philosophical zombies'* (see Kirk, 2019, for a review). The physical facts about such a 'zombie' are just the same as those about an actual human being: The zombie's body, including its brain, is physically indistinguishable from the human's, molecule for molecule. The same functionally characterised psychological states are realised in both creatures and produce the same outputs or effects for the same inputs. Accordingly, the zombie's physical behaviour is exactly the same as the human's. But, unlike the human, the zombie lacks conscious experience – there is nothing it is like to be the zombie, and 'all is dark inside'. This scenario is meant to be *logically* possible – regardless of whether it is consistent with laws of nature (or 'nomologically possible').

Conceivability arguments infer logical possibility from conceivability (e.g., Descartes, 1641; Locke, 1700/1979; Jackson, 1982; Kripke 1972/1980). The *zombie argument* (Chalmers, 1996) assumes that philosophical zombies are conceivable. The simplest version starts out from the conceivability intuition that 'It is conceivable that biological beings with bodies like ours behave just like us and yet do not enjoy any conscious experience.' It further assumes that what is conceivable is logically possible. It infers from these two assumptions that philosophical zombies are logically possible. Both assumptions are controversial. We will empirically examine the conceivability intuition in the first assumption (criticized by, among others, Dennett, 1995; Kirk, 2008; Marcus, 2004; Thomas, 1998; *cf.* Woodling, 2014).

Different versions of the argument employ different notions of conceivability. Stronger notions of conceivability make the conceivability assumption more controversial but the inference from conceivability to possibility more compelling (Kirk, 2019). In response to this challenge, the argument's chief proponent, Chalmers (2002; 2010), elaborated epistemic notions of conceivability potentially strong enough to license the inference to logical possibility: A proposition or statement S is *epistemically conceivable* when it is conceivable that S is actually the case, for all we know a priori (Chalmers, 2002, p.157). Such conceivability can be negative or positive. The latter is the better guide to possibility (Chalmers 2002, p.160):

> S is *positively conceivable* for a thinker when the thinker 'can imagine a coherent situation that verifies S, where a situation verifies S when, under the hypothesis that the situation actually obtains, the [thinker] should conclude that S' (Chalmers, 2010, p.146).

According to Chalmers, conceivability needs to be established in two stages (Chalmers, 2002, pp.147-148): First, a thinker intuitively judges that a statement S passes the relevant test for conceivability, and reflectively endorses this judgment. For *positive conceivability*, this means conducting what we will call the

> POSCON TEST: A thinker tries to imagine a situation that verifies S, hypothetically assumes that this situation is actual (rather than counterfactual), and intuitively assesses whether it follows from that assumption that S is the case (*cf.* Chalmers, 2002, pp.157-158).

For example, the thinker tries to imagine a 'zombie scenario' and assesses whether, if the scenario is actual, it will be the case that there are biological beings with bodies like ours, who behave just like us, but do not enjoy any conscious experience. This test yields intuitions about whether the imagined situation verifies the target statement ('There are biological beings…'). S is '*prima facie* conceivable' if and only if the verification intuitions are positive and are reflectively endorsed. Such reflectively endorsed intuitions provide defeasible evidence for genuine conceivability. In a second stage, theorists examine and exclude potential defeaters. These include philosophical arguments that question the apparent coherence or imaginability of the relevant situation (Chalmers, 2010, pp.154-155) – and debunking explanations of the

initial verification intuition.

In a classic survey of professional philosophers' beliefs, Bourget and Chalmers (2014) used agreement ratings to determine whether philosophers believe that philosophical zombies are conceivable: 16% of respondents judged philosophical zombies inconceivable, about 60% deemed them conceivable. (The remainder declared themselves agnostic or unfamiliar with the literature or found the question too unclear.) While it is unclear how many respondents (if any) rehearsed the POSCON test, these findings provide initial evidence that philosophical zombies are prima facie conceivable for academic philosophers. We now build up to a debunking explanation of conceivability assessments that stands to defeat any evidence provided by verification intuitions or agreement with conceivability claims.

1.3. Linguistic salience bias

To meet specific research needs, philosophers often give new senses to words that already have well-established uses in ordinary discourse. The zombie argument introduces a new sense of 'zombie', now recognised by the *Oxford Dictionary*.[1] As Chalmers explains, the argument applies the word to beings that are identical with us in every physical and behavioural respect, 'quite unlike the zombies to be found in Hollywood movies' (Chalmers, 1996, p.95). The new sense's explanation thus explicitly cancels various implications from the word's dominant 'Hollywood' sense (e.g., that the 'zombies' have rotting bodies, lifeless faces, etc.). But established habits of thought may make it difficult to apply new explanations consistently. Dennett (1995, p.322) suggested that 'when philosophers claim that zombies [in their new sense] are conceivable, they invariably underestimate the task of conception (or imagination), and end up imagining something that violates their own definition.' We will develop and experimentally test a psycholinguistic explanation that lets us understand when and why competent thinkers fail to abide by their own definition of 'zombie'.

When people hear or read utterances, they immediately deploy knowledge about the world, to interpret the utterances (Elman, 2009; 2011). *Stereotypes* are implicit knowledge structures in semantic memory that encode information about statistical regularities observed in the physical or discourse environment (McRae & Jones, 2013). Stereotypes can be associated with individual nouns and verbs (and are then also known as 'prototypes' or 'situation schemas', respectively). As traditionally conceived, such stereotypes represent clusters of weighted features that come to mind first when we encounter those words and are diagnostic or predictive of the relevant categories (Hampton, 2006). As evidenced by priming experiments (review: Lucas, 2000), single words ('tomato') activate stereotypical features (*red*) rapidly (within 250ms) (review: Engelhardt & Ferreira, 2016). Nouns and verbs together ('The mechanic checked…') can swiftly activate complex stereotypes that encode information about recurrent situations (car inspections) and are not activated by individual words on their own (Bicknell et al., 2010; Matsuki et al., 2011).

Activated stereotypes support automatic *stereotypical inferences* to attributions of stereotypical features (the tomato talked about is red). In co-operative communication (Grice 1989), such defeasible inferences are made by hearers and anticipated by speakers (Levinson 2000; *cf.* Garrett & Harnish 2007): Speakers typically skip mention of stereotypical features but make deviations from stereotypes explicit ('the green tomato'). In the absence of such explicit indications to the contrary, hearers assume the situation talked about conforms to the relevant stereotypes, treating the most specific stereotypes activated (say, about car inspections) as the most relevant. Stereotypical inferences that clash with contextual information or background

[1] Sense 1.3 in: https://www.lexico.com/definition/zombie . Last accessed Jan.6, 2021.

knowledge can be suppressed within one second (Fischer & Engelhardt, 2017; *cf.* Faust & Gernsbacher, 1996). Less dramatically, initial activation simply decays in the absence of contextual support (Oden & Spira, 1983). Together with the rapid deployment of the most specific stereotypes, these mechanisms ensure that stereotypical inferences are highly context-sensitive, and that contextually inappropriate inferences hinder comprehension and further reasoning only rarely.

A complication arises, however, from the way many polysemous words are processed (for reviews, see Eddington & Tokowicz, 2015; Vicente, 2018). In general, polysemes activate a unitary representation of semantic information that is deployed to interpret utterances which use the word in different senses (Macgregor, Bouwsema, & Klepousniotou, 2015; Pylkkänen, Llinás, & Murphy, 2006). The findings about how words cue world knowledge for rapid deployment in utterance interpretation (see above) suggest a unitary representation is typically built around stereotypical features associated with the word. Different senses can sometimes be generated by rules (as in metonymy) and sometimes not (as in metaphor). In the latter case, of 'irregular polysemy', the unitary representation consists in overlapping clusters of semantic features (Brocher, Foraker, & Koenig, 2016; Klepousniotou et al., 2012), and may include overlapping stereotypes.

For example, the dominant sense of 'zombie' is made up of the features that contribute to the 'Hollywood' stereotype. This includes the typical features (I) *attacks and eats humans*, *rotting body*, and *infected* as well as (II) *reanimated*, *lacks free will*, and *lifeless face*, (III) *dead inside*, *move slowly*, and *dumb* (see Appendixes A-B) and *lacks conscious experience* (Appendix D). One subordinate sense is associated with the partially overlapping 'voodoo' stereotype. This excludes (I) but includes features from (II) and (III) and adds features *is under a magic spell* and *under others' control*. Another subordinate sense, the metaphorical sense ('Before my morning coffee, I am a zombie'), is not associated with a distinct stereotype but shares features (III) with the dominant sense.

The verbal stimulus activates the features shared by related senses quickly and strongly, regardless of context, while the activation of unshared features is a function of their relative exposure frequency (Brocher et al., 2018): The more often the language user encounters the word in one sense, rather than another, the more strongly the (unshared) features associated with (only) that sense are activated, when the user encounters the word. This baseline activation may be boosted by context (ibid.). Another factor determining strength of activation is prototypicality: Features deemed to make for particularly good examples of the relevant category (say, 'zombie') are activated more rapidly and strongly (Hampton, 2006). Strength of activation thus depends on *linguistic 'salience'*. Unlike the contextual salience involved in familiar salience biases (see Taylor & Fiske, 1978, for a classical review), this is not a contextual magnitude, but a function of relative exposure frequency over time and prototypicality (Giora, 2003).

Due to those processing properties of irregular polysemes, pronounced salience imbalances between their dominant and subordinate senses give rise to bad inferences. When the dominant sense of an irregular polyseme is far more salient than its other senses, the features strongly associated with the dominant sense will be strongly activated by the verbal stimulus, regardless of context. Where the word is used in a subordinate sense which shares some, but not all, of these frequently co-instantiated features with the dominant sense, appropriate interpretation of utterances which use the subordinate sense will require suppression of some of these features, while retaining some others (*cf.* Giora's (2003) Retention/Suppression Hypothesis). To interpret philosophical uses of 'zombie' (which speak of 'zombies' that behave like you and me, and have bodies like ours, but lack conscious experience), we need to suppress

stereotypical features including *attacks and eats humans*, and *has a rotting body*, while retaining the feature *lacks conscious experience*. But frequently co-occurring component features of an activated stereotype exchange lateral cross-activation (Hare et al., 2009; McRae et. al., 2005). Where such cross-activation of contextually irrelevant components complements their initial strong activation due to salience, their complete suppression is impossible. Merely partially suppressed features continue to support inferences – e.g., from 'Fred is my zombie twin' to *Fred attacks humans*.

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

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

then

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

In a nutshell: When encountering unbalanced irregular polysemes whose interpretation involves suppression, thinkers are liable to be swept along by stereotypical inferences, even when these are defeated by the context. Contextually inappropriate but persistent inferences predicted by linguistic salience bias have been documented by studies combining eye tracking (measurements of pupil dilations and reading times) and plausibility ratings (Fischer & Engelhardt 2017; 2019; 2020; Fischer, Engelhardt, & Sytsma, 2020).

1.4. Hypotheses and preliminary studies

We suggest that linguistic salience bias provides a (debunking) explanation of conceivability judgments about philosophical zombies: Automatic comprehension inferences shape our verdicts about verbally described cases: they routinely have us infer what else is also true of the cases described. Such defeasible inferences are then integrated (as 'inferred content'; *cf.* Machery, 2017, p.13) into the situation model, i.e., the mental representation of the situation talked about, on which further judgment and reasoning about the situation are based (Zwaan, 2016). In this way, defeasible pragmatic inferences, including stereotypical inferences (Levinson, 2000), shape philosophical thought experiments (Saint-Germier, 2021) and arguments (Fischer et al., 2021). Three corpus studies, four surveys, and an experiment examined the suggestion that philosophical uses of the irregular polyseme 'zombie' are affected by linguistic salience bias and trigger contextually inappropriate but persistent stereotypical inferences that influence judgments about the conceivability of philosophical zombies.

Preliminary studies (reported in Appendices A-E) established that the three conditions that jointly give rise to the salience bias – (i) to (iii) above – are all met by the word 'zombie' and subordinate senses including its philosophical sense. Three corpus studies (see Appendix A) confirmed that, as per condition (i), the 'Hollywood' sense of zombie is far more salient than all other senses (accounting, e.g. for over 80% of occurrences in a random sample from the *Corpus of Contemporary American English*). A typicality rating task (see Appendix B) elicited features stereotypically associated with the dominant sense of 'zombie', and confirmed that these include several broadly physical and behavioural features (*rotting bodies*, *lifeless face,* etc.), which are cancelled by explanations of the philosophical sense which require that zombies be physico-behaviourally indistinguishable from normal humans. Hence the intended

interpretation of philosophical uses of 'zombie' requires suppression of those cancelled features, as per condition (ii). A plausibility rating task (see Appendix C) provided evidence that when imaginary beings are characterised as 'zombies' people infer these beings lack conscious experience. Two typicality rating studies (reported in Appendices D and E) used positive and negative consciousness attributions, respectively, to provide evidence that the stereotypes associated with the dominant sense of 'zombie' includes the feature *lacks conscious experience*. This feature is also relevant for the philosophical sense, as per condition (iii). Thus, philosophical uses of 'zombie' meet all conditions for linguistic salience bias. We therefore hypothesise that this bias extends to these uses. In other words:

> **H1** Competent language users make contextually cancelled stereotypical inferences from philosophical uses of 'zombie' and these influence their judgment and reasoning about philosophical 'zombies'.

This hypothesis requires experimental investigation: Our plausibility rating task (Appendix C) revealed that people automatically infer possession of conscious experience from the information that the beings at issue are physico-behaviourally indistinguishable from 'normal' humans (*cf.* Arico et al., 2011, on inferences from specific behavioral cues). Explanations of the *philosophical* sense of 'zombie' that convey this information will therefore be perceived as cancelling the stereotypical feature *lacks conscious experience*. This means that a tension is built into the philosophical sense of 'zombie', from the start, as the stereotypical feature *lacks conscious experience* is both contextually relevant and cancelled. Arguably, a similar tension is built into the dominant 'Hollywood' sense: Both *lacks conscious experience* and *feels hungry* are deemed typical of zombies (Appendices B and E), even though the latter would seem to imply conscious experience. This situation – where a stereotypical feature simultaneously is contextually relevant and cancelled (or inconsistent with another contextually relevant feature) – has not been experimentally investigated before.

H1 suggests a debunking explanation of the targeted conceivability intuitions: Linguistic salience bias will tip the balance of activation in favor of the stereotypical feature (*lacks conscious experience*), against the information implied by the cancelling context. This means that if linguistic salience bias applies to 'zombie', it will affect how thinkers assess the positive conceivability of philosophical zombies, with Chalmers's POSCON test (see above, Sect. 1.2): The zombie argument invites us to try to imagine a situation in which scientists, evolution, or Divine intervention have created 'zombies' that possess bodies like ours and behave like us (=**P**) but have no conscious experience (=**~Q**). To assess whether philosophical zombies are prima facie positively conceivable, thinkers should assume that the imagined scenario is actual, and consider whether the relevant statement 'P&~Q' follows from this assumption, so that the imagined situation verifies that statement. As we just noted, people make spontaneous inferences from 'P' to *possesses conscious experience* ('Q') and will normally take a situation that verifies 'P' to verify 'Q', rather than '~Q'. But if salience bias tips the balance of activation in favour of the stereotypical zombie feature '~Q', against the information 'Q' implied by 'P', competent language users will be more willing to accept the situation imagined as verifying both 'P' and '~Q', and hence 'P&~Q'. Linguistic salience bias will thus promote intuitive verification judgments that provide defeasible evidence for the conceivability of philosophical zombies. To put it more succinctly, in Chalmers's (2002, p.147) terms:

> **H2** Linguistic salience bias contributes to rendering philosophical zombies prima facie conceivable.

The finding that the relevant verification judgments are largely due to this bias – and are rarely

made when verbal prompts do not give rise to the bias – would undermine their status as evidence of conceivability.

2. Experiment

Our main study examined **H1** and **H2** and measured the prevalence of conceivability intuitions.

2.1 Predictions

Our hypotheses predict framing effects: In the philosophical sense, 'zombie' is logically equivalent to 'duplicate that is physico-behaviourally indistinguishable but lacks conscious experience'. But 'duplicate' lacks the stereotypical associations of 'zombie' that are cancelled by the requirement of physico-behavioural indistinguishability. **H1** predicts that when vignettes describe imaginary beings as 'zombies', rather than as 'duplicates', participants will be more inclined to accept attributions of typical zombie properties that are cancelled by that requirement. Similarly, **H2** predicts that when participants assess the conceivability of beings that are physico-behaviourally indistinguishable from us but lack conscious experience, these beings will prove *prima facie* conceivable for more participants when described as 'zombies', rather than as 'duplicates'. The most straightforward approach would test this prediction by presenting participants with differently phrased zombie scenarios and asking them to judge the conceivability of the imagined beings ('zombies' or 'duplicates'). However, laypeople may not possess or deploy the relevant concept of conceivability.

A *pre-study* examined this possibility (see Appendix F). To qualify as the genuine article, conceivability judgments need to show a strong negative correlation with judgments of contradictoriness: The more clearly contradictory a scenario is, the less conceivable it is. The pre-study therefore presented participants with a philosophical zombie scenario that used the phrase 'duplicate' (to avoid further complications) and elicited agreement ratings for judgments of conceivability and contradictoriness. The study revealed there was no significant negative correlation (not even trending towards significance). Over a third of participants declared themselves agnostic ('neutral') 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). 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.

We therefore implement Chalmers's POSCON test (Sect. 1.2) to assess prima facie positive conceivability: Our participants read a vignette that uses either the word 'zombie' or 'duplicate' in inviting them to imagine the creation of beings that are physico-behaviourally indistinguishable from us (=**P**) but where 'all is dark inside' (=**~Q**). The vignette thus seeks to prompt participants to imagine philosophical zombies (if they can). The vignette places the scenario in the future of the actual world and encourages participants to consider the scenario they imagine as actual. A subsequent agreement rating task then elicits judgments about what will be true if that scenario is actual, and thus tests to what extent participants take the scenario they imagine to verify attributions of lack of consciousness as well as attributions of typical and atypical zombie properties. Agreement ratings thus provide evidence that the beings they imagine fit descriptions including 'P & ~Q' and qualify as philosophical zombies. In this setting, the hypothesis **H1**, that salience bias extends to philosophical uses of 'zombie', makes a prediction about mean agreement ratings for typical zombie features that are cancelled by the requirement of physico-behavioural indistinguishability and about mean ratings for features

consistent with the requirement but atypical for zombies:

> *[Prediction 1]* Participants will agree more strongly with attributions of typical zombie features (like T1-T3 below) and agree less strongly with attributions of atypical features (like A1-A3 and A-D below), in the Zombie than the Duplicate condition.

The hypothesis **H2**, that salience bias promotes positive verification judgments, concerns a binary choice (agreement vs non-agreement with judgments), rather than strength of agreement. **H2** therefore yields a prediction about proportions (rather than means): more participants in the Zombie than the Duplicate condition will accept the imagined situation as verifying both 'P' and '~Q', where agreement with '~Q' materialises as disagreement with consciousness attributions. I.e.:

> *Prediction 2:* More participants will simultaneously (i) agree that the imagined beings are physico-behaviourally indistinguishable from normal humans *and* (ii) disagree with consciousness attributions (like A-D below), in the Zombie than the Duplicate condition.

2.2. Methods

Participants: Participants were recruited through advertising on Google for a free personality test, which was administered after the main task. While there are notable benefits to employing such a 'push strategy' (see Appendix B), attention tends to be lower than with paid participants (Haug, 2018). We therefore employed an attention check in addition to comprehension checks conceptually required by our research questions (see below). Participants were restricted to native English-speakers raised in North America, 16 years of age or older, with at most minimal training in philosophy. 638 participants met these restrictions. 28.5% of these participants failed the attention check. A further 38.3% failed at least one of the demanding comprehension checks. This left 247 participants.[2] Though substantive, the exclusions did not affect key findings (see Appendix G).

Materials: Each participant read the following vignette, using either the word 'zombie' or 'duplicate':

> Imagine that 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.
>
> Imagine that scientists successfully scan and duplicate an average person in this way. What, if anything, do you think the resulting [zombie/duplicate] would be like?

The vignette includes a literal statement of Chalmers's proposition P and employs a familiar metaphor to state ~Q. P is stated by '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.' Scare quotes around the first occurrence of 'zombie' indicate the word is used here in a special sense, as a synonym of 'exact physical duplicate' in the previous sentence. This implicitly cancels all assumptions of similarity between these 'zombies' and Hollywood zombies. ~Q is stated by the final sentence '… all is dark inside.'

The metaphorical formulation (also used by Chalmers, 1996, p.96) was chosen because subsequent items will ask participants to rate attributions of conscious experience (A-D below).

---

[2] These participants were 73.3% women (1 non-binary), mean age 42.2 years (16-84 years).

Including in the vignette direct negations of these statements would have invited responses based on shallow processing (Ferreira, Bailey, & Ferraro, 2002; Sanford et al., 2006), i.e., without attempting to integrate information from different parts of the vignette. This would have created the risk of subsequent responses to consciousness attributions being based only on statements of ~Q and failing to take into account the information about physico-behavioural indistinguishability (P). This would have prevented us from interpreting responses as evidence for or against **H2**. The metaphorical formulation promotes deeper cognitive engagement with the vignette and facilitates responses to items that take into account all relevant information. This allows us to interpret subsequent responses to consciousness attributions as relevant to **H2** (see Sects. 3.1 and 3.4 for further discussion).

Participants rated agreement with statements using a 7-point scale (anchored at 1 with 'totally disagree', at 4 with 'neither agree nor disagree', and at 7 with 'totally agree'). Statements included four attributions of consciousness:

A The [zombie/duplicate] would be capable of having conscious experiences.
B The [zombie/duplicate] would have an inner mental life, including feelings and emotions.
C The [zombie/duplicate] would sentient and experience its surroundings and sensations.
D There is something it would feel like to be the [zombie/duplicate].

In addition, participants rated three statements attributing clusters of typical zombie features (T1-T3) and three statements attributing clusters of atypical zombie features (A1-A3), obtained through a prior typicality rating study (see Appendix B):

T1 The [zombie/duplicate] would have a rotting body and attack and eat humans.
T2 The [zombie/duplicate] would move slowly and have a lifeless face.
T3 The [zombie/duplicate] would lack free will and feel no joy.

A1 The [zombie/duplicate] would be capable of being sad and feeling hate.
A2 The [zombie/duplicate] would think and be intelligent.
A3 The [zombie/duplicate] would be capable of being happy, singing, smelling flowers, and feeling love.

T1 and T2 are clearly cancelled by the vignette's statement of physico-behavioral indistinguishability (P), T3 arguably so.

Design and procedure: Participants were administered demographic questions, the rating task, and a 'big five' personality test, in this order. Demographic questions included questions concerning educational attainment (highest level completed), level of education in relevant subjects of study (philosophy, psychology and the brain sciences, natural sciences), and religiosity (participation with an organized religion). In the main task, we manipulated a single variable (*term*) with two levels ('zombie' vs 'duplicate'), between subjects. Nine items (T1-T3, A1-A3, consciousness attributions A-C) and an attention check were presented below the vignette on the same page, giving participants the opportunity to consult the vignette before answering. Items were presented in random order. To identify those participants who responded in awareness of the contextual information 'P' that cancels several stereotypical inferences from 'zombie', we added a second page with two comprehension checks:

(1) According to the story, the [zombie/duplicate] has a brain just like the original person.
(2) According to the story the [zombie/duplicate] behaves differently from the original person.

The vignette was not repeated, and participants were unable to return to the previous page. (This and the fact that (1) requires agreement and (2) disagreement rendered these checks quite

demanding.) Since preliminary studies had provided evidence that participants found the 'Nagelian' item D even harder to understand when presented side-by-side with A-C (see Appendix D, esp. Fn.25), we presented this item separately, on the second page. The three items on page two were presented in random order.

To study whether people make contextually cancelled stereotypical inferences that influence further cognition, we need to ensure that participants are aware of the relevant contextual information – here 'P' (the imagined beings have bodies like the original person and behave like that person). Where this information clashes with persistent stereotypical inferences, it may be taken into account through suspension of judgment: In response to the subsequent comprehension questions, participants may choose to neither agree nor disagree that the beings have bodies like ordinary humans or behave like us (rating '4'). By contrast, incorrect disagreement ('1'-'3') with comprehension check (1) and incorrect agreement ('5'-'7') with check (2) manifest lack of awareness of 'P'. We therefore excluded from the analysis participants who failed the attention check, disagreed with the first comprehension check, or agreed with the second. Exclusions left 122 and 125 participants, respectively, in the Zombie and Duplicate conditions.

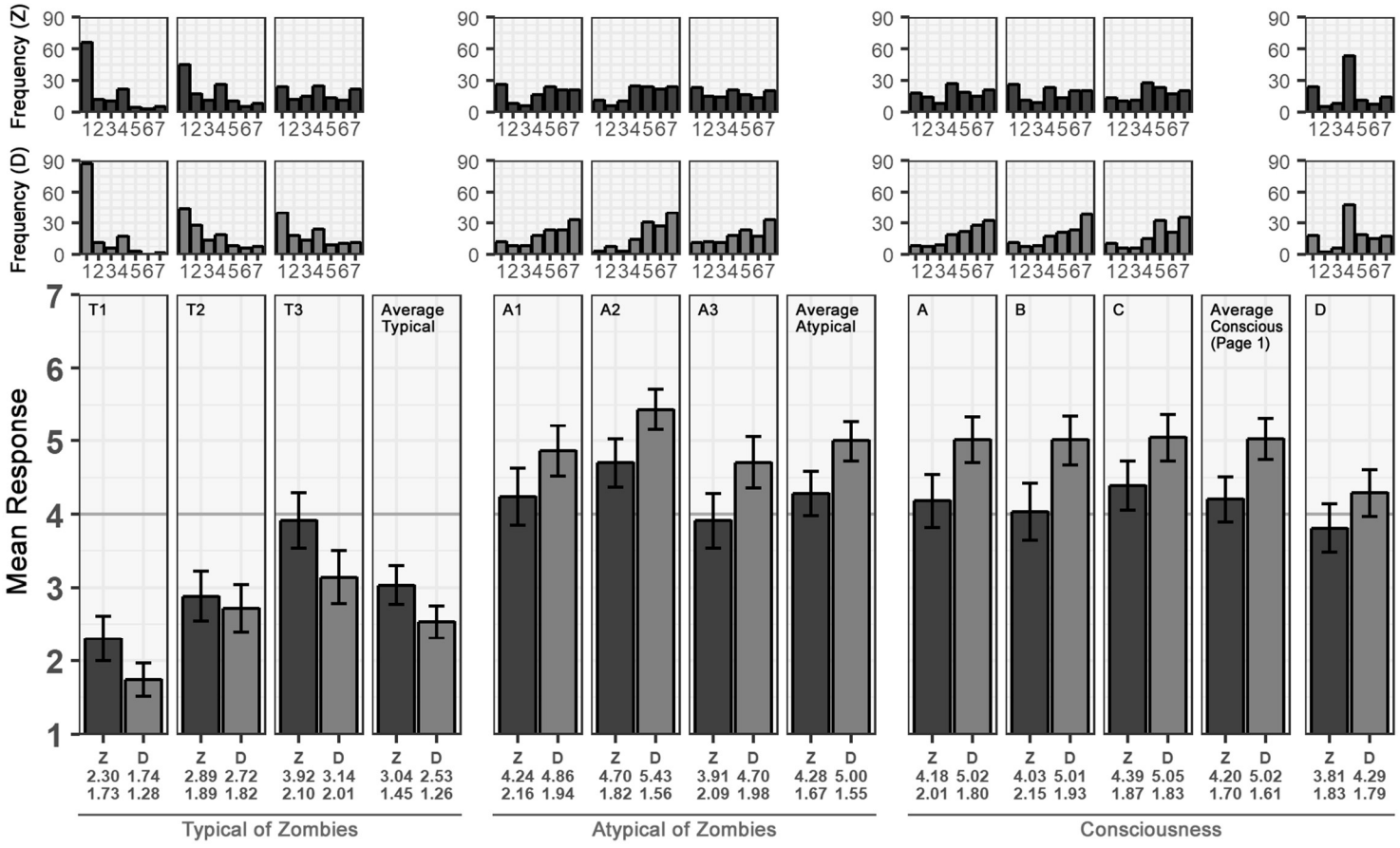

Figure 1. Results for Main Study with means followed by standard deviations below the bar graphs, for zombie (Z) and duplicate (D) conditions; bar graphs showing 95% confidence intervals. Histograms above each bar graph show the frequency distributions of responses across participants for each condition.

2.3. Results and discussion

Findings are presented in Figure 1. Our hypotheses predict an effect of *term* ('zombie' vs 'duplicate') for each of our three types of *cluster* (typical=T1-T3, atypical=A1-A3, consciousness=A-D). We therefore first conducted a two-way mixed ANOVA with *term* as a

between-subjects factor and *cluster* as a within-subjects factor, controlling for variation between participants across the ten items. Responses to T1-T3 were reverse coded, so that all items were treated as attributing properties atypical of zombies, as per previous typicality ratings. This analysis revealed the predicted main effect of *term* $F(1,245)=16.61$, $p<.001$, $\eta^2=.028$ and a main effect of *cluster* $F(2,2219)=52.326$, $p<.001$, $\eta^2=.025$.

To examine **H1** in more detail, follow-up analyses considered item categories individually. These analyses confirmed the predicted effect of *term* for typical items $F(1,245)=8.478$, $p=.0039$, $\eta^2=.017$ and atypical items $F(1,245)=12.29$, $p<.001$, $\eta^2=.033$ (see Appendix G2 for more detail). Next, we examined the consciousness items (A-D). On their intended interpretation, all these items entail possession of conscious experience. Findings from preliminary studies had suggested that participants do not give the intended interpretation to D and had us place this item on page two. We therefore first examined responses to A-C, on page one. To determine whether our participants treated these items similarly, a one-way repeated-measures ANOVA examined ratings for them in the zombie and the duplicate condition separately. There was no effect of *item*, in either condition (zombie: $F(2,242)=2.187$, $p=.12$, $\eta^2=.005$; duplicate: $F(2,248)=.043$, $p=.96$, $\eta^2=.000$). A follow-up reliability analysis carried out on A-C revealed a high Cronbach's alpha ($\alpha=0.83$). These findings confirm that participants treated items A-C similarly. We therefore summed the responses. As predicted, a Welch's t-test revealed that ratings were significantly higher in the duplicate condition than in the zombie condition and the effect size was medium $t(243.41)=3.92$, $p<.001$, $d=.50$, one-tailed. Finally, we even found an effect of term for the remaining Nagelian item D $t(244.5)=2.07$, $p=.020$, $d=.26$, despite evidence that many participants failed to understand this item as intended or at all.[3] Indeed, the term manipulation made a categorial difference to consciousness attributions: While the crucial items A-C were deemed distinctly plausible (mean ratings A-C significantly above '4') in the duplicate condition $t(124)=7.13$, $p<.001$, $d=.64$, two-tailed, they dropped to 'neutral' in the zombie condition $t(121)=1.30$, $p=.20$, $d=.12$, two-tailed.

Finally, the high internal consistency of consciousness items observed across different studies[4] motivated the follow-up question of whether these items capture an underlying folk conception of conscious experience. To address this question, we conducted a latent variable exploratory factor analysis using ordinary least squares to find the minimum residual solution for responses to items A-D. Both parallel analysis and the number of eigenvalues greater than 1 (2.30, 0.96, 0.39, 0.35) indicated a single factor. Factor loadings were high for A-C (0.78, 0.83, 0.77), but low for D (0.17). This indicates that A-C are indeed capturing an underlying latent variable and further justifies our focus on these items in the above analysis.

Present findings confirm *Prediction 1*: Even though relevant stereotypical implications were cancelled by contextual information, attributions of typical zombie properties were accepted more strongly in the Zombie than the Duplicate condition, while attributions of atypical zombie properties, including attributions of conscious experience, were accepted less

---

[3] As in previous typicality rating studies (Appendices D and E), we observed marked spikes of neutral '4' ratings for D, from participants who agreed with A-C, which entail D on its intended Nagelian interpretation. Moreover, including D with A-C in a one-way repeated-measures ANOVA similar to the above, we found an effect of *item* for both conditions (zombie: $F(3,363)=3.003$, $p=.030$, $\eta^2=.011$; duplicate: $F(3,372)=8.049$, $p<.001$, $\eta^2=.029$) not observed for A-C alone.

[4] These are the two typicality rating studies using A-D and negative counterparts, respectively, reported in Appendices D and E, and the main study. Cronbach's alpha for items A-C: $\alpha=0.76$ in first typicality study, $\alpha=0.83$ in main study; for negative counterparts: $\alpha=0.73$. For items A-D: $\alpha=0.72$ in first typicality study, $\alpha=0.73$; for counterparts: $\alpha=0.74$.

in the zombie condition. The observed framing effect provides evidence that linguistic salience bias extends to 'zombie'.

To assess to what extent this bias affects the prima facie positive conceivability of philosophical zombies (as per **H2**), we further restricted our sample and turned from means to frequencies. On Chalmers's approach, philosophical zombies are thus conceivable by participants if and only if participants take a situation they imagine to verify the crucial statements 'P' and '~Q' (Sec.1.2). Participants in our study thus need to assume that the beings imagined are physico-behaviourally indistinguishable from normal humans ('P'). This means that to assess **H2** we need to work with even stricter restrictions than in the study of salience bias: To study whether people make contextually cancelled stereotypical inferences that influence further cognition, we 'merely' had to ensure that participants take into account the relevant contextual information. In our experiment, such awareness of the relevant information 'P' was consistent with suspension of judgment on our two comprehension-check items (see Sec. 2.2). To conceive of philosophical zombies, however, participants need to agree that these beings have bodies like ordinary people and disagree that their behaviour differs. We therefore further restricted attention to those 188 participants who gave these responses ('5'-'7' and '1'-'3', respectively) to our two comprehension-check items (rather than '4'-'7' and '1'-'4', as before). These participants further need to judge that the imagined scenario verifies the statement ('~Q') that the imagined beings lack conscious experience. In our experimental set-up, this verification judgment translates into disagreement with A-D (or at any rate A-C, given the intelligibility issue concerning D). We therefore determined the proportion of these participants (N=188) that disagreed with consciousness attributions A-D (rated them '1'-'3') in the zombie and the duplicate condition, respectively. Findings are reported in Table 1 (below).

Depending upon the formulation (A-D) used, philosophical zombies seem prima facie conceivable, on Chalmers's approach (the POSCON test), for roughly 30-40% of our participants – when the imagined beings are called 'zombies'. Crucially, this figure drops notably, when these beings are described as 'duplicates' – to roughly 15-20% of participants. As predicted by **H2**, the wording thus makes a significant difference to participants' verification judgments (see χ^2-tests for A-C in Table 1). When readily intelligible formulations are used, 38.1% (A), 49.7% (B), and 49.4% (C) of the disagreement with the attribution of conscious experience elicited in the zombie condition disappears in the duplicate condition; calling duplicates 'zombies' makes it almost twice as likely (1.62 times for A, 1.99 times for B, and 1.97 times for C) that people will judge that imagined duplicates lack consciousness (and verify both 'P' *and* '~Q'), when invited to imagine philosophical zombies (physico-behavioural duplicates where 'all is dark inside'). This provides evidence that the framing effect observed in previous analyses affects the verification judgments that provide defeasible evidence for the conceivability of philosophical zombies.

Table 1

Percentages of participants disagreeing with consciousness attributions A-D in zombie and duplicate conditions, while accepting physico-behavioural indistinguishability, with statistical comparisons between conditions.

	A	B	C	D
Zombie (N=93)	32.3	39.8	31.2	31.2
Duplicate (N=95)	20.0	20.0	15.8	22.1
χ^2	3.056	7.875	5.383	1.546
p-value	0.04	0.003	0.01	0.107

Finally, we examined two predictions concerning demographic factors. Recent findings showed that salience bias is not mitigated by higher verbal IQ (Fischer, Engelhardt, & Herbelot, ms). Educational attainment is a reasonable proxy for general and verbal IQ: Educational level and numbers of years in education show substantive correlations with general intelligence and vocabulary (Kaufman, Reynolds, & McLean, 1989; Strenze, 2007; Uttl & Pilkenton-Taylor, 2001). We therefore predicted that educational attainment would not notably affect ratings. This was tested by looking at the Spearman's rho correlation coefficient between summed ratings for attributions A-C and educational level (highest degree) for the maximally restricted sample (N=93 [Z], 95 [D]) using two-tailed tests. We did not find a significant correlation overall (r_s=.0076, p=.85) or for either the duplicate condition (r_s=-.036, p=.73) or the zombie condition (r_s=.083, p=.42) individually. Similarly, no significant correlations were found for extent of training in psychology or the brain sciences (r_s=.022, p=.58; [Z] r_s=-.065, p=.54; [D] r_s=-.079, p=.44) or the sciences more generally (r_s=.049, p=.22; [Z] r_s=.030, p=.78; [D] r_s=.16, p=.13). (To be maximally conservative, no corrections were applied for multiple tests.)

Second, belief in body-soul dualism is highly correlated with religiosity and influences inferences from descriptions of bodily states to attributions of experience (Gray, Knickman, & Wegner, 2011). Religious participants are more likely to regard a soul distinct from the body as the locus of conscious experience and to regard such souls as being imparted by God. We therefore expected them to be less likely to infer that a being created by scientists possesses conscious experiences, whether that being is described as a 'duplicate' or a 'zombie', and predicted negative correlations between religiosity and attributions of consciousness (summed ratings for A-C). Using one-tailed tests, we found the predicted negative correlation overall (r_s=-.15, p<.001), and for the duplicate condition (r_s=-.19, p=.031) and the zombie condition (r_s=-.22, p=.019), individually. Participants self-identifying as more religious were less likely to agree with the consciousness attributions. However, the size of the observed correlations suggests the influence of the relevant cultural beliefs on consciousness attributions was weak.

3. General discussion

3.1 Main findings

We investigated the supposedly widely shared intuitive judgments that provide the key evidence for the positive conceivability of philosophical zombies. In a lay sample free of prior philosophical commitments, we found that these judgments ('verification judgments') display framing effects that are due to linguistic salience bias.

Our main study implemented Chalmers's test procedure for positive conceivability (the POSCON test; see Sect. 1.2) in a way that takes into account the relevant empirical constraints. Participants were tasked with imagining beings that are physico-behaviourally indistinguishable from normal humans but 'all dark inside'. When the beings to be imagined were described as 'zombies' rather than 'duplicates', participants were more inclined to attribute typical zombie features to the imagined beings, and less inclined to attribute atypical zombie features. This was so even though the typical features were inconsistent with the contextual information of physico-behavioural indistinguishability, and the atypical properties were consistent with it. These framing effects are evidence of contextually cancelled stereotypical inferences from philosophical uses of 'zombie' (as per **H1**). Corpus studies and typicality rating studies established that these uses satisfy the conditions triggering linguistic salience bias (Sect. 1.4), which can explain such inappropriate inferences (Sect. 1.3). We infer that the observed framing effects are due to linguistic salience bias. The observed effects were

mainly small, in contrast with the large effect sizes observed in previous studies of the bias (Fischer & Engelhardt, 2019; 2020). This may be due to the internal tension in the philosophical sense of 'zombie' examined (Sect. 1.4; for psycholinguistic discussion, see Appendix H).

We observed a framing effect of medium size for attributions of conscious experience: Markedly more participants were inclined to accept that the imagined beings lack conscious experience when these beings were described as 'zombies' rather than, neutrally, as 'duplicates' (as per **H2**). Depending upon how consciousness attributions were phrased, participants were 1.6 times to twice as inclined when the vignette spoke of 'zombies'. We infer our main conclusion:

(1) In zombie thought experiments using the word 'zombie', linguistic salience bias accounts for up to half of the apparent prima facie positive conceivability of philosophical zombies.

The moment we described philosophical zombies in less tendentious terms, only few participants in our lay sample passed Chalmers's (POSCON) test for positive conceivability:

(2) Depending upon how consciousness attributions were phrased, only 15%-20% of participants made the requisite verification judgments, when the beings to be imagined were described neutrally as 'duplicates'.

Generalisations from this second finding seem open to an objection: To avoid shallow processing (Sect. 2.2), we used only the metaphor 'all is dark inside' to prompt participants to imagine beings that lack conscious experience (~Q). But our plausibility study (Appendix C) revealed that 'all is dark inside' has an interpretation compatible with possession of conscious experience ('full of bad thoughts and feelings'). Our vignette offers no contextual support for this interpretation (there is no suggestion that the 'average person' getting duplicated is so negative). But other formulations could still have prompted more participants to strain their imagination and come up with scenarios they would have been willing to accept as verifying both 'P' *and* '~Q'.

The only related findings available speak against this suggestion: In a study on modal intuitions, Peressini (2014, pp.874-5) used literal formulations to ask participants about the possibility of a 'medical procedure that would remove your inner experience without affecting your brain, so from the outside you would remain unchanged physically and behaviourally' and the possibility of 'a person physically and behaviourally identical to you in all ways but who had no inner experience at all'. Proportions for the (undergraduate) sample (8% and 12%) were yet lower than the relevant proportions in our study (Peressini, personal communication). This may be due to the framing ('you' and 'person' vs our neutral 'duplicate').

Our plausibility rating study confirmed that laypeople make inferences from physico-behavioural indistinguishability (P) to possession of conscious experience (Q) (see Appendix C). In the absence of pragmatic factors relevant only where vignettes explicitly refer to conversational contexts, participants will therefore endorse inferences from lack of conscious experience (~Q) to lack of indistinguishability (~P), i.e., the presence of some physico-behavioural difference (Bonnefon & Villejoubert, 2007). More effective prompting to imagine a scenario verifying '~Q' therefore stands to prevent more participants from agreeing that the requirement of physico-behavioural indistinguishability (P) is met. Comparison with Peressini's (2014) related findings suggests our vignette was as effective as possible at prompting participants to imagine a being participants are prepared to accept as sporting both key features of philosophical zombies. We conclude that for a majority of our lay participants philosophical zombies are not prima facie positively conceivable.

3.2 Assessing the zombie argument

These findings help us assess the zombie argument (Sect. 1.2). The zombie argument proceeds from the premise that philosophical zombies are conceivable and infers that such beings are possible. Something is conceivable if its prima facie conceivability is not defeated or 'undermined' by further reflection or findings. Hence, 'to reject the premise, one needs to find something that undermines the prima facie … imaginability' (Chalmers, 2010, p.154). We found (1) linguistic salience bias and (2) majority dissent, in a study of positive conceivability, which is the best guide to possibility (Chalmers, 2002, p.160). As we now argue, both findings provide undermining defeaters (Pollock, 1985) that defeat the evidence for conceivability provided by what positive verification judgments we have observed.

Linguistic salience bias renders thinkers unable to fully suppress stereotypical inferences that are cancelled by contextual information – or which they would otherwise take to be cancelled by contextual information – and thus prevents them from taking contextual information into account in the way they would when unaffected by the bias. This neglect of contextual information renders the resulting judgments epistemically deficient and deprives thinkers of warrant for accepting them (for a review of relevant philosophical debate, see Machery, 2017, pp.90-125). In the present case, stereotypical inferences from 'zombie' to *lacks conscious experience* are deemed inconsistent with the contextual information that the 'zombies' at issue are physico-behaviourally indistinguishable from normal humans, as evidenced by the inferences from physico-behavioural normality to conscious experience observed in our plausibility rating study (Appendix C). Where thought experiments use the word 'zombie', salience bias prevents thinkers from taking this contextual information fully into account when assessing whether the 'zombies' they imagine verify not only 'P' but also '~Q' ('The imagined being lacks conscious experience'). Finding (1) implies that this failure to take essential contextual information sufficiently into account accounts for up to half of the positive verification judgments. The latter provide defeasible evidence that philosophical zombies are ideally positively conceivable. Finding (1) thus provides an undercutting defeater (Pollock, 1985) that undermines this evidence, where it is provided by thought experiments which use the word 'zombie'.

When considering prima facie conceivability as defeasible evidence for conceivability, we should hence take into account only verification judgments elicited in thought experiments that use more neutral terms like 'duplicate'. In such a thought experiment, only 15-20% of laypeople found philosophical zombies prima facie positively conceivable (Finding 2). But thought experimentation is not democratic: a bright few may get right what the many get wrong. Our analysis of demographic factors (end Sect.2.3) helps assess whether the 15-20% minority was in a better epistemic position, or possessed better conceptual competence, to address the difficult task. We found that religiosity weakly correlated with increased prima facie conceivability. Training in natural science, or in psychology, did not correlate with consciousness attributions. Nor did educational attainment, which can serve as a proxy for verbal and general IQ (Kaufman, Reynolds, & McLean, 1989; Strenze, 2007; Uttl & Pilkenton-Taylor, 2001). We tentatively infer that the minority was not in a better epistemic position than the majority, nor did they benefit from greater conceptual competence. We accordingly treat this as a case of peer disagreement.

The leading epistemological positions on peer disagreement then lead to the same ultimate conclusion (*cf.* Machery, 2017, pp.131-136): Participants making divergent verification judgments in the zombie thought experiment will base their endorsement of conflicting intuitive judgments on little, and mostly the same, evidence – mainly features mentioned in the brief case description. According to the 'Total Evidence view' (e.g., Kelly,

2011) what it is reasonable to believe depends on both the original, first-order evidence and on 'higher-order evidence' afforded by 'the fact that one's peers believe as they do' (p.201). Where peers arrive at the relevant belief independently (as in our experiment with lay participants), this higher-order evidence is additive (p.205). Where an endorsement is supported by insubstantial first-order evidence, but an overwhelming majority of peers independently arrives at a dissenting judgment, this higher-order evidence will therefore 'swamp the first-order evidence into virtual insignificance' (p.203) and render it rational to reject the initial minority judgment. The 'Equal Weight view' (e.g., Elga, 2007) accords each peer's view equal weight in determining what level of credence to give a judgment. Since a large majority disagrees with it, the minority judgment should be given a lot less credence than the majority judgment. On either approach, the evidence for conceivability provided by the minority's verification judgments is rendered insignificant by their disagreement with the large majority. The Finding (2) of 'majority dissent' undermines evidence for the positive conceivability of philosophical zombies, also where this is provided by thought experiments that avoid the word 'zombie' (as in our duplicate condition). If our vignette was maximally effective (Sect. 3.1), we can infer that philosophical zombies are not positively conceivable for laypersons.

Our first finding, of salience bias, also undermines what evidence of positive conceivability is provided by survey responses of expert philosophers. In Bourget and Chalmers's (2014) survey of professional philosophers, 60% of respondents deemed philosophical zombies conceivable, in a question format that was ambiguous between negative and positive conceivability and explicitly referred to 'zombies'. Against reasonable expectations, also this indirect evidence is undermined by linguistic salience bias: Higher verbal IQ (Carroll, 1993; Deary et al., 2007) and executive functioning (Miyake & Friedman, 2012) can be reasonably expected to shield philosophers' reflective judgment from being affected by salience bias. Moreover, differences in linguistic diet may reduce the salience imbalances to which they are exposed, in the first place – e.g., philosophers may encounter 'zombie' more frequently in its philosophical sense and less frequently in its Hollywood sense than laypeople. Finally, philosophers' higher reflectivity, as documented with the Cognitive Reflection Test (Livengood et al., 2010), might prevent them from basing further judgment or reasoning on contextually cancelled default inferences they cannot help making. Even so, framing effects predicted by the salience bias hypothesis have been replicated with professional philosophers, including philosophers exposed to specialist discourse in which salience patterns for words of interest are flipped by comparison with general discourse (Fischer, Engelhardt, & Herbelot, ms). These findings suggest that salience bias affects also expert philosophers' judgments and reasoning about philosophical zombies, including their survey responses.

While some philosophers, including Chalmers (2018), hold that the philosophical notion of 'phenomenal consciousness' captures a folk-psychological concept, empirical studies have provided a more nuanced picture (Peressini, 2014) and evidence to the contrary (Knobe & Prinz, 2008; Sytsma & Machery, 2010; Sytsma, 2016; Sytsma & Ozdemir, 2019). The factor analysis of responses to our consciousness items (Sect. 2.3) added to this evidence: Loadings on the single factor were high for A-C – and very low for the Nagelian formulation D that philosophers use to explain 'phenomenal consciousness'. This suggests there is a live risk that philosophical concepts ('phenomenal consciousness', 'qualia', 'functional role', etc.) and theories do not make it possible to conceive of duplicates that lack 'conscious experience' in the folk sense but, rather, create cognitive artifacts (*cf.* Machery, 2017, p.90).

To sum up: The two findings of our main study, (1) linguistic salience bias and (2) majority dissent, suggest that philosophical zombies are not positively conceivable for philosophically untrained laypeople and undermine extant evidence for the positive

conceivability of philosophical zombies, provided by lay *or* expert responses. But such conceivability is required to secure the zombie argument's inference from conceivability to possibility. So far, this argument does not get off the ground.

3.3 Implications for the hard problem of consciousness

The zombie argument is meant to justify the so-called 'modal intuitions' that are central to the philosophical hard problem of consciousness (Sect. 1.1): In the standard dialectic (Chalmers, 1996, pp.93-108), proponents of the argument invoke these modal claims to justify acceptance of 'metaphysical intuitions' that assert conscious phenomena are distinct from physical events ('The feeling of pain is a different thing than neural activity in the DPI') and 'explanatory intuitions' that assert physical facts cannot explain phenomenal facts ('Neural activity in the DPI cannot explain what it is like to feel pain'). Together, these 'problem intuitions' suggest that *any* scientific explanation of how physical processes give rise to conscious experience is bound to fall short of what is required (Block, 1995; Sytsma 2009).

The prevalence of explanatory and metaphysical problem intuitions has recently begun to be studied: Gottlieb and Lombrozo (2018, Study 3A) had participants recruited through M-Turk rate their agreement with claims of the form 'Science could one day fully explain the following phenomenon' about various psychological phenomena. Problem intuitions are indicated by disagreement with claims about paradigmatic conscious experiences like 'having headaches' (7% of participants) and 'discerning temperature through touch' (8%). Low disagreement suggests few participants thought science is in principle barred from explaining the kind of conscious experiences at the heart of the philosophical debate. Diaz (forthcoming) had participants recruited through M-Turk rate their agreement with explanatory claims like 'The properties of pain are fully explained in terms of neural activity in the DPI' and metaphysical claims like 'The feeling of pain is just neural activity in the DPI'. Diaz's three studies used different claims. Their higher specificity might account for higher disagreement (20-25%) than in Gottlieb & Lombrozo (2018). Even so, none of these problem intuitions was shared by more than a quarter of participants, in any study.

This low prevalence of problem intuitions means that the philosophical hard problem of consciousness is not foisted on us by widely shared folk beliefs or 'common sense'. Rather, the impression that scientific explanations of conscious experience are bound to fall short arises from philosophical arguments. The philosophical hard problem of consciousness can therefore be 'dissolved' by refuting those arguments and thereby showing that the impression of a principled obstacle to scientific explanation is unwarranted. We refuted the zombie argument that is central to the standard dialectic, by showing that the conceivability assumption it starts from is unwarranted, as it rests on epistemically deficient intuitions.

Our findings support the suggestion that no one mechanism or factor explains all problem intuitions or even solely accounts for any one class of problem intuition (*cf.* Dennett, 2019). We identified two factors accounting for conceivability intuitions: Linguistic salience bias besetting a comprehension inferences, which can account for up to half of the examined positive conceivability intuitions about 'zombies'. Religious belief further contributes to such intuitions. The identification of a bias affords a debunking explanation. Religious belief has been repeatedly targeted by naturalistic debunking explanations (for reviews see Leben, 2014; Leech & Visala, 2011). Our findings thus contribute to efforts to 'dissolve' the philosophical hard problem by developing complementary debunking explanations of problem intuitions, for each class of such intuitions, that reveal we have no warrant for accepting those intuitions (*cf.* Arico et al., 2011; Fiala & Nichols, 2019; Graziano, 2019; Webb & Graziano, 2015). The increasingly rich literature on the 'meta-problem of consciousness' suggests a variety of partially competing,

partially complementary cognitive, linguistic, cultural, historical, and sociological factors that may be implicated in generating one or more classes of problem intuitions (for a review, see Chalmers, 2020a; 2020b).

Further development of these explanations will let us assess the key contention suggested by present findings: The impression that there is a *principled* obstacle to scientific explanations of how physical processes give rise to conscious experience is generated by philosophical arguments that ultimately rely on epistemically deficient intuitions.

3.4 Future research

The present study initiated the investigation of conceivability intuitions elicited in philosophical thought experiments. It identified several constraints facing future research on such intuitions. Our pre-study suggests that conceivability intuitions cannot be elicited from philosophically untrained participants, by asking whether scenarios are conceivable or contradictory. Our main study trialled an alternative approach that implements Chalmers's (2002) test for positive conceivability: A vignette prompts participants to imagine a scenario of interest and to consider it as actual. Subsequent agreement ratings elicit judgments about whether the scenario verifies the statement of interest. Given the nature of philosophical conceivability questions, this statement will typically state that features (like our *P* and *Q*) that typically go together do not do so in the scenario (e.g., that 'normal' human bodily behaviour does not go with conscious experience). To describe the scenario of interest, it is often hard to avoid giving new or special senses to polysemous words, whose interpretation requires suppression of stereotypical features associated with the word. This gives rise to linguistic salience bias.

To be able to produce evidence of positive conceivability, the vignette must therefore eschew subordinate uses of polysemes, in particular of words whose stereotypical implications would influence the assessment of the target statement ('P and ~Q'). Second, the vignette must be effective at prompting participants to imagine a scenario that fits the brief (i.e., verifies both 'P' and '~Q'). Third, vignette and items must be phrased or presented in a way that prevents shallow processing, i.e., ensures that participants take all the relevant information (both 'P' and '~Q') into account, when making their verification judgments. The last two requirements are in tension: Including in the vignette, for effectiveness, explicit statements of 'P' and '~Q' risks verification judgments 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 shallow processing risks to leave participants without sufficiently clear guidance for the imagination task. The present study negotiated these constraints by using the neutral 'duplicate' and resorting to metaphor ('all is dark inside' for the relevant instance of '~Q'). We would welcome future research, for a start on the conceivability of philosophical zombies, that explores different ways of negotiating these constraints.

3.5 Conclusion

The linguistic salience bias (Fischer & Engelhardt, 2019; 2020) is poised to arise when philosophers give a special sense to words (irregular polysemes) that already have a related, but distinct dominant sense from ordinary discourse – as philosophers frequently do (Fischer et al., 2021). The resulting unwarranted inferences are then bound to lead to wrong conclusions when thinkers use the subordinate sense to talk about cases that pull apart features that usually go together. Many philosophical thought experiments envisage such cases (Machery 2017, pp.116-18). Linguistic salience bias is thus set to vitiate philosophical thought experiments of this kind when they employ unbalanced irregular polysemes in subordinate senses. The bias results in framing effects and generates epistemically deficient zombie intuitions: intuitions that are

'killed' (defeated) by contextual information but kept cognitively alive by the bias. This paper documented framing effects. The observed effects show the bias affects judgment and reasoning in philosophical thought experiments that use the word 'zombie': It contributes to generating conceivability intuitions about philosophical zombies that few people share when cases are described in neutral terms. Such intuitions about zombies are zombie intuitions. Their exposure as zombie intuitions may contribute to 'dissolving' the philosophical hard problem of consciousness: to debunking the impression that *principled* obstacles prevent scientific explanations of how physical processes give rise to conscious experience, and showing that this impression rests on epistemically deficient intuitions.

Acknowledgements
For helpful comments on previous drafts, we thank the editor and three anonymous reviewers for this journal as well as David Chalmers, Rodrigo Diaz, Paul E. Engelhardt, Eyuphan Ozdemir, Kevin Reuter, and audiences at three conferences: the 3rd Conference of the Experimental Philosophy Group Germany, the Corpus Fortnight of the Australasian Experimental Philosophy Group, and the 94th Joint Session of the Aristotelian Society and Mind Association. Dan Simpsonbeck and Esther Marshall annotated corpora in Study 1C. This research was supported by internal grants from the University of East Anglia and the Victoria University of Wellington.

Author credit statement
Eugen Fischer: Conceptualisation, Methodology, Writing – Original Draft. Justin Sytsma: Methodology, Formal Analysis, Investigation, Data Curation, Writing – Review and Editing. Both authors contributed equally to the design of the study and the interpretation of results.

Bibliography
Arico, A., Fiala, B., Goldberg, R. & Nichols, S. (2011) The folk psychology of consciousness. *Mind and Language*, 26, 327–352.

Bicknell, K., Elman, J.L., Hare, M., McRae, K., & Kutas, M. (2010). Effects of event knowledge in processing verbal arguments. *Journal of Memory and Language, 63,* 489–505.

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences,* 18, 227-287.

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

Bonnefon, J.-F., & Villejoubert, G. (2007). Modus Tollens, Modus Shmollens: Contrapositive reasoning and the pragmatics of negation. *Thinking & Reasoning*, 13(2), 207-222

Brocher, A., Foraker, S., & Koenig, J.-P. (2016). Processing of irregular polysemes in sentence reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*(11), 1798–1813

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

Byrd, R. J., Calzolari, N., Chodorow, M.S., Klavans, J.L., Neff, M.S. and Rizk, O.A. (1987). Tools and methods for computational lexicology, *Computational Linguistics,* 13, 219-40.

Carroll, J. B. (1993). *Human Cognitive Abilities: A Survey of Factoranalytic Studies*. New York: Cambridge University Press.

Chalmers, D.J. (1996). *The Conscious Mind*. New York: Oxford University Press.

Chalmers, D.J. (2002). Does conceivability entail possibility? In T. Gendler, & J. Hawthorne (eds.), *Conceivability and possibility* (pp. 145-200). Oxford: Oxford University Press

Chalmers, D.J. (2010). The Two-Dimensional Argument Against Materialism. In his: *The Character of Consciousness* (pp. 141-191). Oxford: Oxford University Press.

Chalmers, D.J. (2018). The meta-problem of consciousness. *Journal of Consciousness Studies*, 25 (9-10), 6–61

Chalmers, D.J. (2020a). How can we solve the meta-problem of consciousness? *Journal of Consciousness Studies*, 27(5-6), 201-226

Chalmers, D.J. (2020b). Is the hard problem of consciousness universal? *Journal of Consciousness Studies*, 27(5-6), 227-257.

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

Deary, I.J., Strand, S., Smith, P., & Fernandes, C. (2017). Intelligence and educational achievement. *Intelligence*, 35, 13–21.

Dennett, D. (1995). The Unimagined Preposterousness of Zombies', *Journal of Consciousness Studies*, 2: 322–6.

Dennett, D.C. (2019). Welcome to strong illusionism. *Journal of Consciousness Studies*, 26(9-10), 48-58.

Descartes, R. (1641). Meditations on First Philosophy. In Cottingham, J., Stoothoff, R., & Murdoch, D. (Eds.), *The Philosophical Writings of René Descartes, Vol. II* (pp.1-62). Cambridge: CUP (1984).

Diaz, R. (forthcoming). Do people think consciousness poses a hard problem? Empirical evidence on the meta-problem of consciousness. *Journal of Consciousness Studies*

Eddington, C.M., & Tokowicz, N. (2015). How meaning similarity influences ambiguous word processing: the current state of the literature. *Psychonomic Bulletin & Review*, *22*(1), 13–37.

Elga, A. (2007). Reflection and disagreement. *Nous*, 41, 478-502.

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

Elman J.L. (2011). Lexical knowledge without a lexicon? *The Mental Lexicon*, *6*(1), 1–33.

Engelhardt, P.E., & Ferreira, F. (2016). Reaching sentence and reference meaning. In Knoeferle, P., Pyykkonen, P., & Crocker, M.W. (Eds.), *Visually situated language comprehension* (pp. 127-150). Amsterdam: John Benjamins

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

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

Fiala, B. & Nichols, S. 2019. Generating explanatory gaps. *Journal of Consciousness Studies*, 26 (9-10), 71-82.

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

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

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

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

Fischer, E., Engelhardt, P.E., Horvath, J., & Ohtani, H. (2021). Experimental ordinary language philosophy: A cross-linguistic study of defeasible default inferences. *Synthese,* 198, 1029–1070.

Fischer, E., Engelhardt, P.E. & Sytsma, J. (2020). Inappropriate stereotypical inferences? An adversarial collaboration in experimental ordinary language philosophy. *Synthese*. 2020; 1-42. https://doi.org/10.1007/s11229-020-02708-x

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

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

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

Gray, K., Knickman, A., & Wegner, D.M. (2011). More dead than dead: Perceptions of persons in the persistent vegetative state. *Cognition*,121, 275–280.

Graziano, M.S.A. (2019). Attributing awareness to others: The attention schema theory and its relationship to behavioral prediction. *Journal of Consciousness Studies*, 26(3-4), 17-37.

Grice, H.P. (1989). Logic and conversation. In his: *Studies in the Ways of Words* (pp. 22-40). Cambridge, Mass.: Harvard UP.

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

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

Haug, M. (2018). Fast, cheap, and unethical? The interplay of morality and methodology in crowdsourced survey research. *Review of Philosophy and Psychology*, 9(2): 363-379.

Jackson, F., 1982, Epiphenomenal Qualia, *Philosophical Quarterly*, 32, 127–136.

Kaufman, A.S., Reynolds, C.R., & McLean, J.E. (1989). Age and WAIS–R intelligence in a national sample of adults in the 20- to 74-year age range: A cross-sectional analysis with educational level controlled. *Intelligence, 13,* 235–253.

Kelly, T. (2011). Peer Disagreement and Higher Order Evidence. In A. Goldman & D. Whitcomb (eds.), *Social Epistemology: Essential Readings* (pp.183-217). Oxford: OUP.

Kirk, R. (2008). The inconceivability of zombies. *Philosophical Studies*, 139, 73–89.

Kirk, R. (2019). Zombies. In E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2019 Edition). https://plato.stanford.edu/archives/spr2019/entries/zombies/

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

Knobe, J. & Prinz, J. (2008). Intuitions about consciousness: Experimental studies. *Phenomenology and the Cognitive Sciences*, 7, 67-85.

Kripke, S. (1972/80), *Naming and Necessity*. Oxford: Blackwell.

Leben, D. (2014). When psychology undermines beliefs. *Philosophical Psychology*, 27, 328–

350.

Leech, D. & Visala, A. (2011). Naturalistic explanation for religious belief. *Philosophy Compass*, 6/8, 552–563

Levine, J. (1983). Materialism and qualia: the explanatory gap. *Pacific Philosophical Quarterly*, 64, 354-361.

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

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

Lucas, M. (2000). Semantic priming without association: a meta-analytic review. *Psychonomic Bulletin and Review*, *7,* 618–630.

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

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

Mallon, R. (2016). Experimental philosophy. In H. Cappelen, T. Szabo Gendler, & J. Hawthorne (eds.), *Oxford Handbook of Philosophical Methodology* (pp. 410-433). OUP.

Marcus, E. (2004). Why zombies are inconceivable. *Australasian Journal of Philosophy*, 82, 477–90.

Matsuki, K., Chow, T., Hare, M., Elman, J.L., Scheepers, C., & McRae, K. (2011). Event-based plausibility immediately influences on-line language comprehension. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *37,* 913–934.

McRae, K., Hare, M., Elman, J.L., & Ferretti, T.R. (2005). A basis for generating expectancies for verbs from nouns. *Memory & Cognition*, *33,* 1174-1184.

McRae, K., & Jones, M. (2013). Semantic memory. In D. Reisberg (ed.), *Oxford Handbook of Cognitive Psychology*, Oxford: OUP.

Miyake, A., & Friedman, N.P. (2012). The nature and organization of individual differences in executive functions: Four general conclusions. *Current Directions in Psychological Science*, 21, 8–14.

Nagel, T. (1974). What is it like to be a bat? *Philosophical Review*, 83, 435–450.

Oden, G.C., & Spira, J.L. (1983). Influence of context on the activation and selection of ambiguous word senses. *Quarterly Journal of Experimental Psychology, 35A,* 51–64.

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

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

Rosch, E., & Mervis, C.B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7, 573-605.

Saint-Germier, P. (2021). Getting Gettier straight: Thought experiments, deviant realization, and pragmatic enrichment. *Synthese*, 198, 1783–1806.

Sanford, A.J.S., Sanford, A.J., Molle, J., & Emmott, C. (2006). Shallow processing and attention capture in written and spoken discourse. *Discourse Processes*, 42, 109-130.

Sękowski, K., Rorot, W. (2021). Intuition-driven navigation of the hard problem of conscious-

ness. *Review of Philosophy and Psychology*. https://doi.org/10.1007/s13164-021-00533-w

Strenze T. (2007). Intelligence and socioeconomic success: a meta-analytic review of longitudinal research. *Intelligence,* 35, 401–426.

Sytsma, J. (2009). Phenomenological obviousness and the new science of consciousness. *Philosophy of Science*, 76, 958-969

Sytsma, J. (2016). Attributions of consciousness. In J. Sytsma & W. Buckwalter (eds.), *A Companion to Experimental Philosophy*, Oxford: Blackwell.

Sytsma, J. & Machery, E. (2010). Two conceptions of subjective experience, *Philosophical Studies*, 151 (2), 299–327.

Sytsma, J. & Ozdemir, E. (2019). No problem: Evidence that the concept of phenomenal consciousness is not widespread. *Journal of Consciousness Studies*, 26 (9-10), 241-256.

Taylor, S.E., & Fiske, S.T. (1978). Salience, attention, and attribution: Top of the head phenomena. *Advances in Experimental Social Psychology*, 11, 249-288.

Thomas, N.J.T. (1998). Zombie killer. In S.R. Hameroff, A.W. Kaszniak, & A.C. Scott (eds.), *Toward a Science of Consciousness II.* (pp. 171–177). Cambridge, MA: MIT Press

Uttl, B., & Pilkenton-Taylor, C. (2001). Letter cancellation performance across the adult life span. *The Clinical Neuropsychologist, 15,* 521–530.

Vicente, A. (2018). Polysemy and word meaning: an account of lexical meaning for different kinds of content words. *Philosophical Studies*, 175, 947-968.

Webb, T.W., & Graziano, M.S.A. (2015). The attention schema theory: a mechanistic account of subjective awareness. *Frontiers in Psychology*, 6:500. doi: 10.3389/fpsyg.2015.00500

Woodling, C. (2014). Imagining zombies. *Disputatio*, 6, 107–116.

Wu, W. (2018). The neuroscience of consciousness. In E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2018 Edition). https://plato.stanford.edu/entries/consciousness-neuroscience/

Zwaan, R.A. (2016). Situation models, mental simulations, and abstract concepts in discourse comprehension. *Psychonomic Bulletin & Review*, 23, 1028–1034.

Supplementary Materials

Appendix A – Corpus studies (Studies 1A-1C)

Study 1A considered collocates in the *Corpus of Contemporary American English* (COCA), starting with the top 20 noun lemmas that occur most frequently directly after 'zombie'.[5] These terms are highly suggestive of Hollywood zombies, with highly used phrases including 'zombie apocalypse', 'zombie movie', 'zombie attack', and 'zombie plague'. This includes nouns like 'mode' and 'walk' that are not immediately suggestive of zombies, but that indicate the Hollywood sense in context: many popular video games have a 'zombie mode' where you fight the walking dead and a 'zombie walk' is a public gathering where people dress up as Hollywood zombies and walk through public spaces. We next looked at the most frequent lemmas to occur within four words of 'zombie'. As seen in Table A-1, these include nouns like 'apocalypse' and 'movie', adjective like 'mindless' and 'flesh-eating', and verbs like 'bite' and 'attack'. These findings suggest the Hollywood sense of 'zombie' is very prominent.

Table A-1

Top 20 nouns, adjectives, and verbs, within four words of 'zombie' in COCA.

Nouns	Count	Adjectives	Count	Verbs	Count
apocalypse	373	Mindless	47	Bite	64
Movie	212	flesh-eating	35	Attack	59
Film	172	Scary	27	Chase	21
Vampire	107	Outer	22	Flick	18
Survival	93	post-apocalyptic	15	Crawl	14
Horde	88	Nazi	14	Overrun	14
Brain	75	Pandemic	13	Shamble	14
Outbreak	61	Millennial	13	Slay	13
Human	52	Creepy	12	Infect	12
Guide	48	Oral	11	Stab	9
Killer	47	Philosophical	11	Roam	9
Mode	44	Resident	10	Outrun	8
Horror	42	Infected	7	Waltz	7
Dragon	42	real-life	7	Unleash	6
Plague	42	full-on	6	Swarm	6
Virus	41	Brainless	6	Lurch	6
Jesus	38	Brainwashed	6	Geeks	6
Novel	37	mixed-up	5	Chomp	6
Invasion	35	Supernatural	5	Undead	6
Genre	34	Apocalyptic	5	Moan	5

Study 1B examined the dominant sense of 'zombie' as revealed by a distributional semantic model (DSM) built from COCA, excluding academic sources. We employed a context-

[5] Top 20 noun lemmas are: 'apocalypse' (count: 364), 'movie' (333), 'film (121), 'survival' (71), 'attack' (57), 'game' (54), 'outbreak' (53), 'horde' (51), 'story' (55), 'plague' (33), 'war' (32), 'mode' (29), 'killer' (29), 'invasion' (28), 'book' (28), 'walk' (27), 'virus' (25), 'bank' (23), 'novel' (21), 'army' (21).

predicting model generated using the word2vec algorithm with standard parameters and a five-word context window. This model has been previously shown to score extremely well on the MEN benchmark with a Spearman's rho of 0.80 (see Sytsma et al., 2019, for details on DSM and the specific models used). The basic idea behind DSMs follows Firth's dictum that 'you shall know a word by the company it keeps' (Firth, 1957, p.11). DSMs look at the company that each word in a corpus keeps—that is, the terms that occur in proximity to it—and assumes that terms that keep similar company have similar meanings. This information is used to represent terms as vectors in a semantic space. The terms can then be compared to give a measure of similarity of meaning, typically by taking the cosine of the vectors. For context-predicting models, this information is used to train an artificial neural network to set vector weights. These models outperform more traditional DSMs (Baroni et al., 2014).

We looked at the nearest neighbours of 'zombie' in the DSM (the terms with the largest cosine values with it and, hence, the terms that are closest in meaning to it according to the model). The nearest neighbours are suggestive of Hollywood monsters, as seen in Table A-2, indicating that the Hollywood sense of 'zombie' is the dominant sense.

Table A-2

Nearest neighbours to 'zombie' in the non-academic COCA DSM created by Sytsma et al. (2019), showing comparison term and cosine value.

Term	Cosine	Term	Cosine
'undead'	0.67	'crazed'	0.55
'vampire'	0.64	'slayer'	0.55
'ghoul'	0.62	'bloodsucking'	0.55
'Dracula'	0.59	'deranged'	0.53
'buffy'	0.58	'ghost'	0.53

Study 1C assessed the linguistic salience of different senses of 'zombie' more directly. Salience is a function of exposure frequency and prototypicality. We used occurrence frequencies in COCA as proxy for exposure frequency and production frequencies as measure of prototypicality. To assess occurrence frequencies, we used a random sample of 500 sentences with the word 'zombie' drawn from COCA. To assess prototypicality, we used a production task and recruited 50 participants using the same method and restrictions as in our main studies.[6] Participants were asked 'Write down 10 sentences that use the word "zombie" in the spaces below. Please try to give varied responses!' We thus obtained 500 produced 'zombie'-sentences. Two independent coders assessed the sentences from the two corpora. The coders were MA students, native speakers of English, ignorant of our research questions. They classified occurrences of the noun 'zombie' as uses of one of 10 senses; where they felt unable to do so, they either indicated insufficient context, that 'zombie' was mentioned rather than used, or that it was ambiguous between Hollywood and voodoo sense (1 or 2 below). The 10 senses were compiled by starting with senses attested by *WordNet*, then adding any further senses and information from *Oxford Dictionaries Online, Oxford English Dictionary*, and *Webster's New World College Dictionary* (see Table A-3).[7]

[6] These participants were 76.0% women, mean age 44.2 years (16-78 years).

[7] https://www.lexico.com/, https://www.oed.com/, https://www.merriam-webster.com/, last accessed 25/7/2019.

Table A-3

Coder evaluations for classifiable sentences in Studies 1C and 1D.

Sense	COCA (1C)			Produced (1D)		
	C1	C2	AVG	C1	C2	AVG
1. 'a corpse reanimated by a virus or other pathogen, typically capable of movement but not of rational thought, and feeding on human flesh, *according to popular fiction and horror movies'*	82.4%	79.8%	81.1%	82.3%	82.3%	82.3%
2. 'a dead body that has been brought back to life, by a supernatural force, *according to voodoo religion*'	0.3%	0.3%	0.3%	0%	0%	0%
3. 'a spirit or supernatural force that reanimates a dead body, *according to voodoo religion or stories*'	0%	0%	0%	0%	0%	0%
4. 'a snake god worshipped by voodoo cults of African origin'	0%	0%	0%	0%	0%	0%
5. 'a person who is listless or lethargic and acts or responds in a mechanical or apathetic way or is dull, slow-witted'.	12.9%	16.1%	14.5%	15.3%	16.1%	15.7%
6. 'a weird, eccentric, or unattractive person', or general term of disparagement	0.9%	0%	0.4%	1.6%	0.4%	1.0%
7. 'a drink made with several kinds of rum, liqueur, and fruit juice'	2.3%	2.6%	2.5%	0.8%	0.8%	0.8%
8. 'i*n philosophy*, a hypothetical being that responds to stimuli as a normal person would but that does not experience consciousness'	0.3%	0.3%	0.3%	0%	0.4%	0.2%
9. Canadian 'man conscripted for home defence' (*Canadian Military slang* 1939–45)	0%	0%	0%	0%	0%	0%
10. 'a computer controlled by another person without the owner's knowledge and used for sending spam or other illegal or illicit activities'	0.9%	0.9%	0.9%	0%	0%	0%

Appendix B – Typicality ratings (Study 2A)

Study 2A used a typicality rating task to identify several features stereotypically associated with 'zombie'.

Participants: As in each study in this paper, participants were recruited through advertising on Google for a free personality test, which was administered after the main tasks.[8] Participants were restricted to native English-speakers raised in North America, 16 years of age

[8] Such 'push strategies' (recruiting participants not directly looking to participate in research) ensure participants are more 'experimentally naïve' and less motivated to provide what they think are experimenters' intended responses (Haug, 2018). Samples collected with the present strategy have been previously compared against samples collected with other methods in replication studies. The present strategy has been consistently found to generate a diverse sample in terms of geography, socio-economic status, religiosity, political orientation, age, and education (e.g., Livengood et al., 2010; Sytsma, 2010; 2012; Sytsma & Ozdemir, 2019).

or older, with at most minimal training in philosophy.[9] 108 participants met these restrictions, with 26.9% of these failing the attention check. This left 79 participants.[10]

Methods: Each participant read 32 sentences attributing a feature to zombies (see Table B) and rated how typical each of these features are for zombies, on a 1-7 scale anchored at 1 with 'very atypical', at 4 with 'neutral (neither typical nor atypical)', and at 7 with 'very typical'. Material development was supported by a norming study, where sixteen psychology undergraduate students, all native speakers of English, first listed typical properties of zombies and then rated the typicality of features the experimenters had suggested based on dictionary explanations and their own intuitions. Based on the norming study, we included 18 presumptively typical zombie features and 14 presumptively atypical features (including two diagnostic of voodoo sense zombies). An attention check was included with the items: 'Please select 5 for this item'.[11] All items appeared in random order.

Results and discussion: To test the difference between the 18 tested features expected to be typical of zombies and the 14 features expected to be atypical of zombies, we began by running a repeated-measures ANOVA with *typicality* as a within-subjects factor, controlling for variation between participants across the 32 items. There was a significant effect for *typicality* and the effect size was large $F(1,2448)=1424$, $p<.001$, $\eta^2=.35$. As seen in Table B-1, the mean for each of the typical features was numerically above the mid-point, while the mean for each of the atypical features was below mid-point. Follow-up tests revealed that this difference was significant for 17 of 18 typical features and for 13 of 14 atypical features. Details for these tests are given in Table B.[12]

Further analyses were conducted to facilitate the assessment of further inferences from feature attributions. Feature attributions work together in promoting further inferences. 'Attacks and eats humans' promotes inferences to 'lion' and 'shark' – and these animals enjoy conscious experience. But 'attacks and eats humans and has a rotting body' may promote different inferences. As a preparatory step for the study of further inferences from attributions of typical zombie features, we therefore need to identify clusters of features that will be coactivated by 'zombie' and will jointly support onward inferences. While rigorous study of coactivation requires priming studies (Lucas, 2000), we reasoned that features that individuals tend to treat as similarly typical or similarly atypical will receive equal amounts of default activation or inhibition from the noun 'zombie' and will, *mutatis mutandis*, influence onward inferences to the same extent.

[9] Minimal training in philosophy was taken to exclude philosophy majors, those who have completed a degree with a major in philosophy, and those who have taken graduate-level courses in philosophy.

[10] These participants were 78.4% women (1 non-binary), mean age 40.9 years (16-67 years).

[11] 26.9% of participants failed the attention check. All the studies reported in this paper used a similar attention check, with just the number participants were asked to select varying.

[12] Throughout we use Student's t-tests for one-sample and paired-sample comparisons and Welch's t-tests for independent-sample comparisons. Since most predictions are directional, we use one-tailed tests unless specified otherwise (like here). We report nonparametric tests—either Wilcoxon signed-rank tests, *W*, or Wilcoxon rank-sum tests, *V*—in parentheses. We also used these tests to confirm the parametric tests reported in the main paper.

Table B

Typical and atypical features from Study 2A in order of descending mean rating, showing mean (SD) after the item and one-tailed comparisons to mid-point below.

Typical Features	Atypical Features
Zombies have rotting bodies. 6.14 (1.53) $t(78)=12.47$, $p<.001$, $d=1.40$ ($V=2549.5$, $p<.001$, $r=0.79$)	**Zombies are alert to their surroundings.** 3.90 (2.08) $t(78)=.43$, $p=.33$, $d=.049$ ($V=831.5$, $p=.27$, $r=.036$)
Zombies attack humans. 6.06 (1.65) $t(78)=11.11$, $p<.001$, $d=1.25$ ($V=2670$, $p<.001$, $r=.75$)	**Zombies feel thirsty.** 3.08 (1.75) $t(78)=4.69$, $p<.001$, $d=.53$ ($V=214.5$, $p<.001$, $r=.46$)
Zombies eat humans. 5.99 (1.67) $t(78)=10.59$, $p<.001$, $d=1.19$ ($V=2476.5$, $p<.001$, $r=.74$)	**Zombies are sad.** 2.86 (1.74) $t(78)=5.83$, $p<.001$, $d=.66$ ($V=144$, $p<.001$, $r=.54$)
Zombies move stiffly. 5.52 (1.92) $t(78)=7.03$, $p<.001$, $d=.79$ ($V=2047$, $p<.001$, $r=.61$)	**Zombies are under a magic spell.** 2.73 (1.84) $t(78)=6.12$, $p<.001$, $d=.69$ ($V=179$, $p<.001$, $r=.56$)
Zombies feel no joy. 5.52 (1.94) $t(78)=6.96$, $p<.001$, $d=.78$ ($V=1969.5$, $p<.001$, $r=.62$)	**Zombies feel hate.** 2.71 (1.63) $t(78)=7.06$, $p<.001$, $d=.79$ ($V=110.5$, $p<.001$, $r=.62$)
Zombies have been infected. 5.49 (2.01) $t(78)=6.60$, $p<.001$, $d=.74$ ($V=2175$, $p<.001$, $r=.60$)	**Zombies think.** 2.70 (1.67) $t(78)=6.92$, $p<.001$, $d=.78$ ($V=133$, $p<.001$, $r=.61$)
Zombies are dead inside. 5.49 (2.02) $t(78)=6.58$, $p<.001$, $d=0.74$ ($V=1969.5$, $p<.001$, $r=.59$)	**Zombies are under others' control.** 2.65 (1.71) $t(78)=7.04$, $p<.001$, $d=.79$ ($V=159$, $p<.001$, $r=.62$)
Zombies feel hungry. 5.42 (1.98) $t(78)=6.35$, $p<.001$, $d=.71$ ($V=1905$, $p<.001$, $r=.58$)	**Zombies are intelligent.** 2.61 (1.44) $t(78)=8.57$, $p<.002$, $d=.96$ ($V=41$, $p<.001$, $r=.70$)
Zombies have lifeless faces. 5.39 (1.91) $t(78)=6.48$, $p<.001$, $d=.73$ ($V=2060.5$, $p<.001$, $r=.59$)	**Zombies are alive.** 2.48 (1.69) $t(78)=8.01$, $p<.001$, $d=.90$ ($V=97$, $p<.001$, $r=.66$)
Zombies move slowly. 5.29 (1.85) $t(78)=6.19$, $p<.001$, $d=.70$ ($V=1572.5$, $p<.001$, $r=.58$)	**Zombies are happy.** 2.13 (1.59) $t(78)=10.48$, $p<.001$, $d=1.18$ ($V=97$, $p<.001$, $r=.76$)
Zombies have a rigid stare. 5.14 (2.00) $t(78)=5.05$, $p<.001$, $d=.57$ ($V=1639$, $p<.001$, $r=.50$)	**Zombies smell flowers.** 2.08 (1.53) $t(78)=11.21$, $p<.001$, $d=1.26$ ($V=61$, $p<.001$, $r=.78$)
Zombies lack free will. 4.95 (2.02) $t(78)=4.18$, $p<.001$, $d=.47$ ($V=1456$, $p<.001$, $r=.42$)	**Zombies talk.** 2.03 (1.54) $t(78)=11.37$, $p<.001$, $d=1.28$ ($V=125$, $p<.001$, $r=.78$)
Zombies feel no pain. 4.91 (2.06) $t(78)=3.92$, $p<.001$, $d=.44$ ($V=1356$, $p<.001$, $r=.41$)	**Zombies feel love.** 1.77 (1.27) $t(78)=15.59$, $p<.001$, $d=1.75$ ($V=13$, $p<.001$, $r=.86$)
Zombies smell blood. 4.91 (1.93) $t(78)=4.20$, $p<.001$, $d=.47$ ($V=1315$, $p<.001$, $r=.45$)	**Zombies sing.** 1.51 (0.96) $t(78)=23.11$, $p<.001$, $d=2.60$ ($V=0$, $p<.001$, $r=.90$)
Zombies have no feelings. 4.89 (2.17) $t(78)=3.63$, $p<.001$, $d=.41$ ($V=1371.5$, $p<.001$, $r=.38$)	
Zombies have been reanimated. 4.85 (1.84) $t(78)=4.10$, $p<.001$, $d=.46$ ($V=1025.5$, $p<.001$, $r=.44$)	
Zombies are dumb. 4.70 (1.60) $t(78)=3.88$, $p<.001$, $d=0.44$ ($V=612$, $p<.001$, $r=.39$)	
Zombies have no moods. 4.34 (2.21) $t(78)=1.37$, $p=.087$, $d=.15$ ($V=1046$, $p=.063$, $r=.16$)	

To identify clusters of features that are deemed similarly typical of zombies, we performed a hierarchical cluster analysis using the Ward algorithm with the Euclidian distance measure. Hierarchical clustering aims to group together items that are similar in the dimension measured by the data it is applied to (here: typicality ratings). The algorithm builds a hierarchy of clusters from the bottom up: each item starts in its own cluster, then the closest two are

combined, with the process repeating until all items are in one cluster. To determine the similarity of single-item clusters, each participant's response for the item is treated as a feature of it, with these values locating the item in an n-dimensional space where n is the number of features (in this case 79, one for each participant in the study). Items that are closer together in this space are treated as more similar. As items are combined into multi-item clusters, the Ward algorithm combines them by minimizing the sum of squares. Applying this procedure for Study 2A produced the cluster dendrogram is shown in Figure B. In line with the previous analysis, we get two high-level clusters, with the 18 typical items clustering together on the right and the 14 atypical items clustering together on the left. We cut the dendrogram at height 18 to give six multi-item clusters – three clusters of typical items (T1-T3) and three clusters of atypical items (A1-A3). The least similar item from cluster A3 ("talk") was removed to reduce this cluster to four items, bringing it more in line with the others (2-3 items).

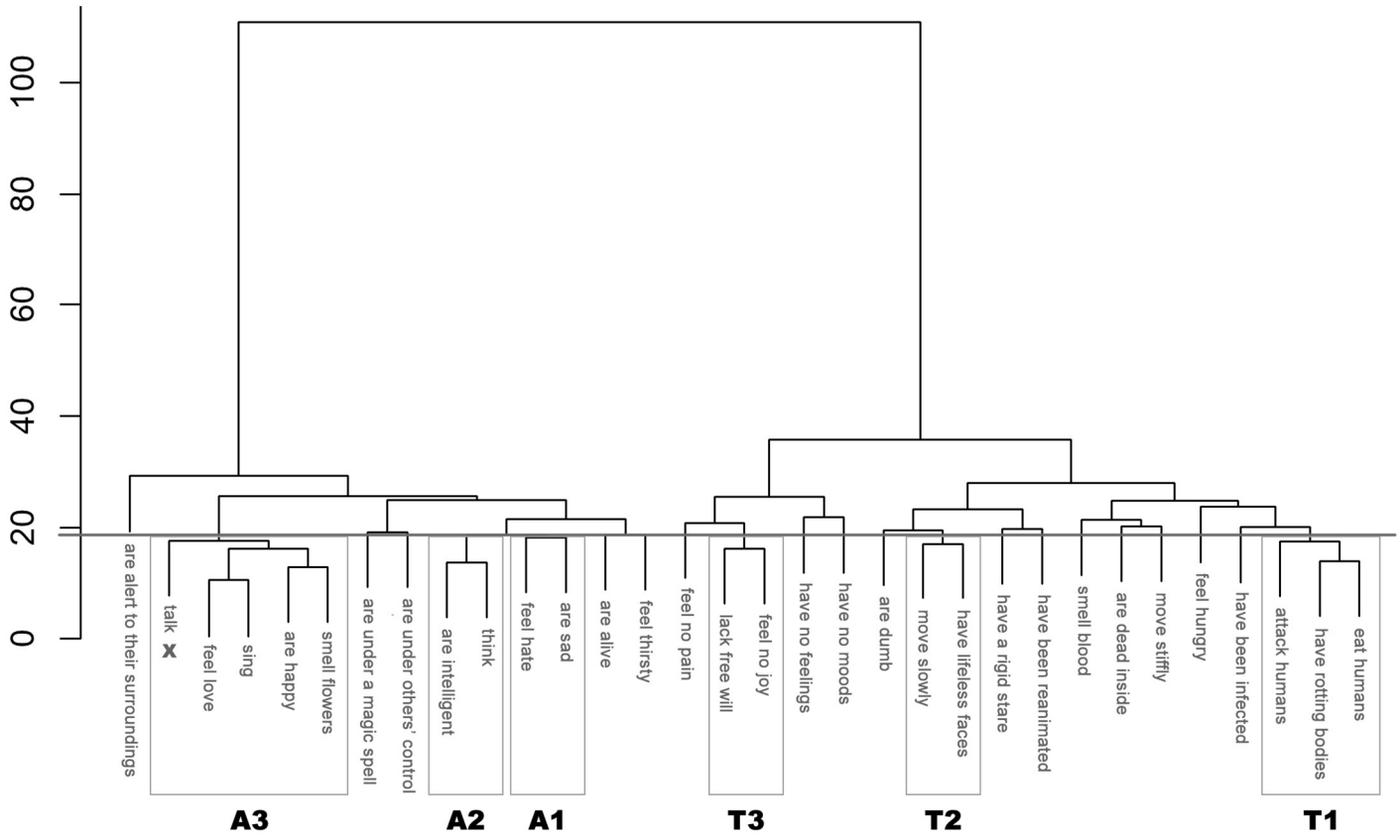

Figure B. Cluster dendrogram for Study 2A, showing cut and resulting clusters used for subsequent studies.

Appendix C – Plausibility ratings (Study 2B)

Study 2B employed a plausibility rating task to examine inferences from attributions of (clusters of) features that are, respectively, typical or atypical for Hollywood zombies, or typical for philosophical zombies. The study examined inferences to attributions of conscious experience (or its lack).

Participants: Participants were recruited in the same way, and with the same restrictions, as in Study 2A. 453 participants met these restrictions, with 34.7% failing one or both of the attention checks and a further 28.0% failing one or both of the comprehension checks described

below. This left 213 participants.[13]

Methods: The study included two pages of critical items. On the first page, participants were invited to imagine 'biological beings' with different sets of properties. For each type of being, critical items then asserted that they enjoy conscious experience, using one of four different formulations: [14]

(A) these beings are capable of having conscious experiences;
(B) these beings have an inner mental life, including feelings and emotions;
(C) these beings are sentient and experience their surroundings and sensations;
(D) there is something it is like to be such a being.

In a between-subject design, each participant saw a version of the questionnaire with one of these four formulations. Participants were instructed: 'The following items invite you to imagine biological beings with certain properties. They then claim that [A/B/C/D]. How plausible is this claim in each case?' Participants were then asked to rate each of eight types of beings. Six were characterized as bearers of one of the six feature clusters T1-T3 and A1-A3 (see Figure B-1) The remaining two types were

(1) 'beings that have bodies like ours and behave like us' and
(2) 'beings that are alive but where everything is dark inside'.

For each of the eight types of being participants were given an item like (example for T1, version A):

> 'Imagine beings that have rotting bodies and attack and eat humans. These beings are capable of having conscious experiences.'

For each item, participants rated the plausibility of the claim on a 1-7 scale anchored at 1 with 'very implausible', at 4 with 'neutral (neither plausible nor implausible)', and at 7 with 'very plausible'. An attention check was included with the items on each of the two pages. All items appeared in random order.

The second page was designed to examine how all features activated by the noun 'zombie', taken together, influence attributions of consciousness. As before, participants were asked to rate the plausibility of consciousness attributions to different kinds of beings. This time, however, the beings were picked out by nouns: 'zombies', 'robots', 'humans', 'dolphins', 'geraniums', 'elves', and 'rocks'. For instance, for 'zombies' the item for consciousness attribution (A) read:

> 'Imagine Zombies. Zombies are capable of having conscious experiences.'

This page used the same 7-point scale as the previous page and included a similar attention check. Items again appeared in random order.

The seven nouns were chosen to give a range of contrasts, including their level of cognitive ability, whether they were fictional, and whether they were alive. 'Humans' was expected to anchor the high-end of the scale, while 'rocks' was expected to anchor the low end. These two items were used as engagement and comprehension checks: Participants who rated

---

[13] These participants were 70.9% women (3 non-binary), mean age 41.4 years (16-77 years).

[14] These formulations have been taken to entail or imply attributions of phenomenal consciousness (Chalmers, 1996, p.xi). Philosophers including Chalmers (2018) hold that this notion captures a folk-psychological concept. Empirical studies have provided a more nuanced picture (Peressini, 2014) and evidence to the contrary (Sytsma & Machery, 2010; Sytsma, 2012; 2016; Sytsma & Ozdemir, 2019). For present purposes we only assume items A-D capture features of mental lives that are directly relevant to the philosophical concept of conscious experience.

consciousness attributions to humans at or below mid-point or attributions to rocks at or above mid-point were excluded from further analysis, for any page. Upon exclusion, a similar number of participants remained in each condition: A (N=57), B (N=51), C (N=54), and D (N=51).

Results and discussion: Findings for the first page are shown in Figure C-1. To test whether the four formulations A-D elicit similar responses across items, we conducted a two-way mixed ANOVA, with *formulation* (A-D) as a between-subjects factor and *item* as a within-subjects factor. For the first page, the analysis revealed a main effect of *item* $F(7,1463)=114.22$, $p<.001$, $\eta^2=.27$, but no main effect of *formulation* $F(3,209)=1.56$, $p=.20$, $\eta^2=.005$, nor an interaction $F(21,1463)=1.24$, $p=.20$, $\eta^2=.009$. That is, ratings were not notably affected by the specific formulation used. We therefore combined conditions A-D for our analyses.

Follow-up tests for the first page indicate that participants tended to find it highly plausible to attribute consciousness (whatever way we phrased it) to (1) beings that 'have bodies like ours and behave like us', with ratings significantly different from mid-point $t(212)=14.42$, $p<.001$, $d=.99$, two-tailed ($V=16961$, $p<.001$, $r=.67$). Attributing consciousness to (2) beings that 'are alive but where all is dark inside' struck participants as less plausible $t(212)=5.75$, $p<.001$, $d=.39$, two-tailed ($V=7005.5$, $p<.001$, $r=.31$). Strikingly, however, participants still tended to rate (2) as distinctly plausible, as assessed against the mid-point $t(212)=6.23$, $p<.001$, $d=.43$, two-tailed ($V=12153$, $p<.001$, $r=.39$). We infer that, in the absence of informative context, 'all is dark inside' has an interpretation compatible with possession of conscious experience ('full of bad thoughts and feelings').

Consciousness attributions to bearers of our atypical zombie property clusters (A1-A3) were all deemed highly plausible, with ratings significantly above mid-point.[15] Interestingly, consciousness attributions to bearers of our typical zombie property clusters (T1-T3) varied notably in their plausibility. While ratings for each of the three clusters differed significantly from mid-point, the mean for T1 was *below* midpoint while the means for T2 and T3 were *above* mid-point.[16] In other words, participants tended to find attributions of conscious experience implausible for T1 but plausible for T2 and T3 (although significantly less plausible for T2 and T3 than for any of A1-A3).[17]

This finding is striking: The behavioural features included in T2 ('move slowly and have lifeless faces'), as well as the properties in T3 ('lack free will and feel no joy'), are intuitively suggestive of diminished conscious experience – and indeed, like T1, attract consciousness ratings that are lower than for beings that (1) 'have bodies like ours and behave like us'.[18] But, in T1, the component 'attack and eat humans' does *not* suggest diminished conscious experience and *cancels* the inference from the remaining component 'have rotting bodies' to the conclusion 'is dead', which would imply 'lacks conscious experience'. So precisely the feature cluster that is least suggestive of lack of conscious experience is the only one that supports inferences to this lack. This finding suggests that T1 attributions support these inferences because this feature cluster is diagnostic of zombies: Participants plausibly infer

---

[15] [A1] $t(212)=15.42$, $p<.001$, $d=1.06$ ($V=17254$, $p<.001$, $r=.72$); [A2] $t(212)=21.20$, $p<.001$, $d=1.45$ ($V=19072$, $p<.001$, $r=.79$); [A3] $t(212)=25.57$, $p<.001$, $d=1.75$ ($V=20363$, $p<.001$, $r=.84$).

[16] [T1] $t(212)=9.25$, $p<.001$, $d=.63$ ($V=3534.5$, $p<.001$, $r=.53$); [T2] $t(212)=4.63$, $p<.001$, $d=.32$, two-tailed ($V=11002$, $p<.001$, $r=.30$); [T3] $t(212)=3.88$, $p<.001$, $d=.27$, two-tailed ($V=11208$, $p<.001$, $r=.26$).

[17] [T2 vs A1] $t(212)=7.74$, $p<.001$, $d=.53$ ($V=1204.5$, $p<.001$, $r=.41$); [T2 vs A2] $t(212)=9.97$, $p<.001$, $d=.68$ ($V=914.5$, $p<.001$, $r=.49$); [T2 vs A3] $t(212)=10.25$, $p<.001$, $d=.70$ ($V=975.5$, $p<.001$, $r=.50$); [T3 vs A1] $t(212)=7.73$, $p<.001$, $d=.53$ ($V=1520$, $p<.001$, $r=.41$); [T3 vs A2] $t(212)=10.41$, $p<.001$, $d=.69$ ($V=959$, $p<.001$, $r=.49$); [T3 vs A3] $t(212)=11.1$, $p<.001$, $d=.76$ ($V=670$, $p<.001$, $r=.53$).

[18] [T1] $t(212)=17.36$, $p<.001$, $d=1.19$ ($V=15123$, $p<.001$, $r=.62$); [T2] $t(212)=6.71$, $p<.001$, $d=.46$, two-tailed ($V=7356.5$, $p<.001$, $r=.35$; [T3] $t(212)=7.34$, $p<.001$, $d=.50$, two-tailed ($V=8820.5$, $p<.001$, $r=.38$).

from T1 features that the beings in question are zombies and infer from this description that they lack conscious experience. The findings from page two speak directly to this suggestion.

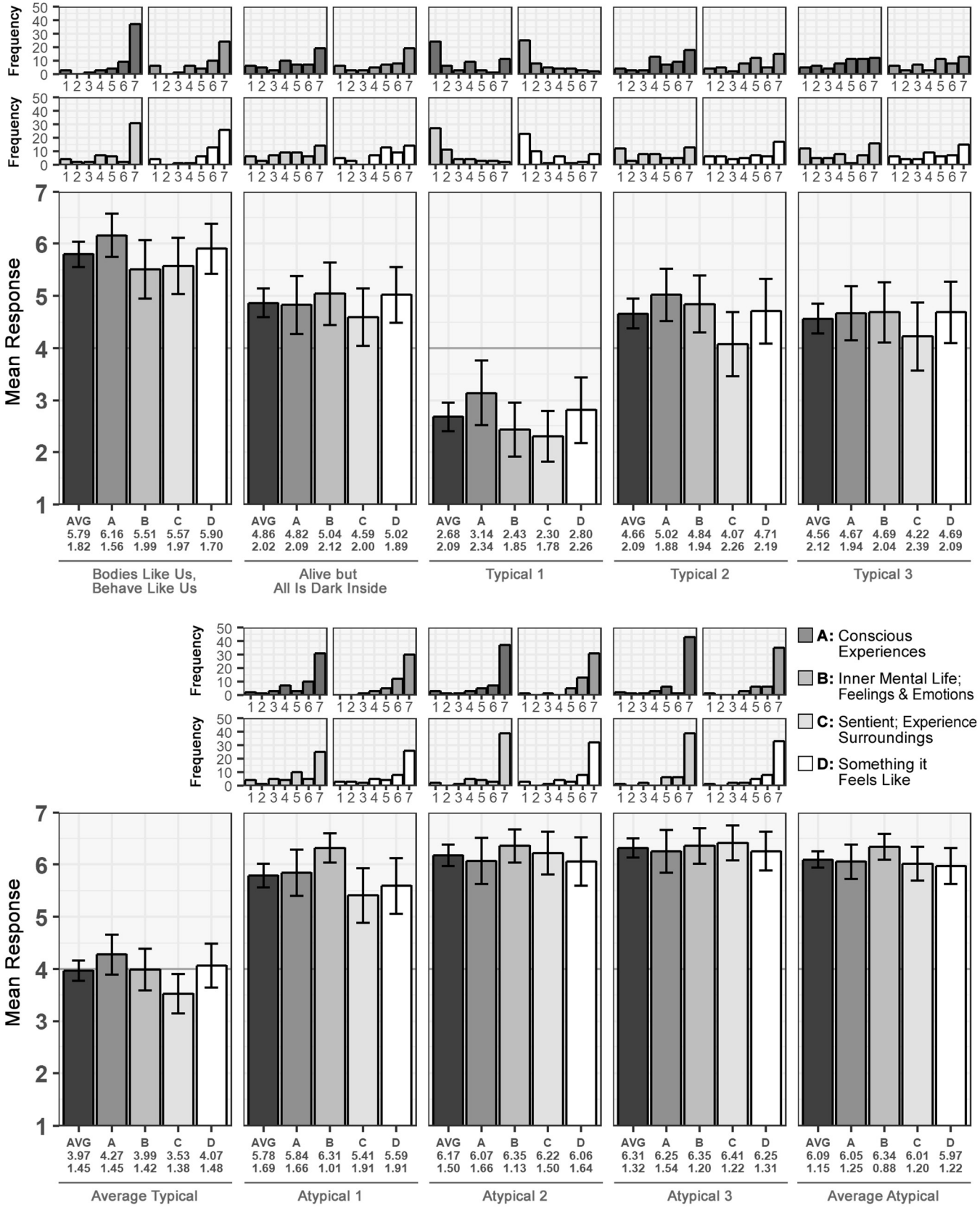

Figure C-1: Results for Study 2-B, Page 1 with means followed by standard deviations below the bar graphs; bar graphs showing 95% confidence intervals. Histograms above each bar graph show the frequency distributions of responses across participants for each condition.

Findings for the second page are presented in Figure 2. A two-way mixed ANOVA for page two again revealed a main effect of *item* $F(4,836)=148.27$, $p<.001$, $\eta^2=.31$ and no main effect for *formulation* $F(3,209)=.51$, $p=.68$, $\eta^2=.001$, although there was now a significant interaction $F(12,836)=7.39$, $p<.001$, $\eta^2=.047$. Consciousness attributions to zombies, however, were similar across the four formulations, with a one-way ANOVA showing no effect for *formulation* $F(3,209)=1.00$, $p=.39$, $\eta^2=.014$. Combining formulations, consciousness attributions to zombies were deemed distinctly implausible (significantly below mid-point) $t(212)=15.25$, $p<.001$, $d=1.05$ ($V=1442$, $p<.001$, $r=.72$). Crucially, consciousness attributions to zombies (on page 2) attracted significantly lower ratings than consciousness attributions to bearers of any of the three clusters of typical zombie features (on page 1), including beings that (as per T1) 'have rotting bodies and attack and eat humans' $t(212)=3.16$, $p<.001$, $d=.22$ ($V=1982.5$, $p=.0014$, $r=.19$).

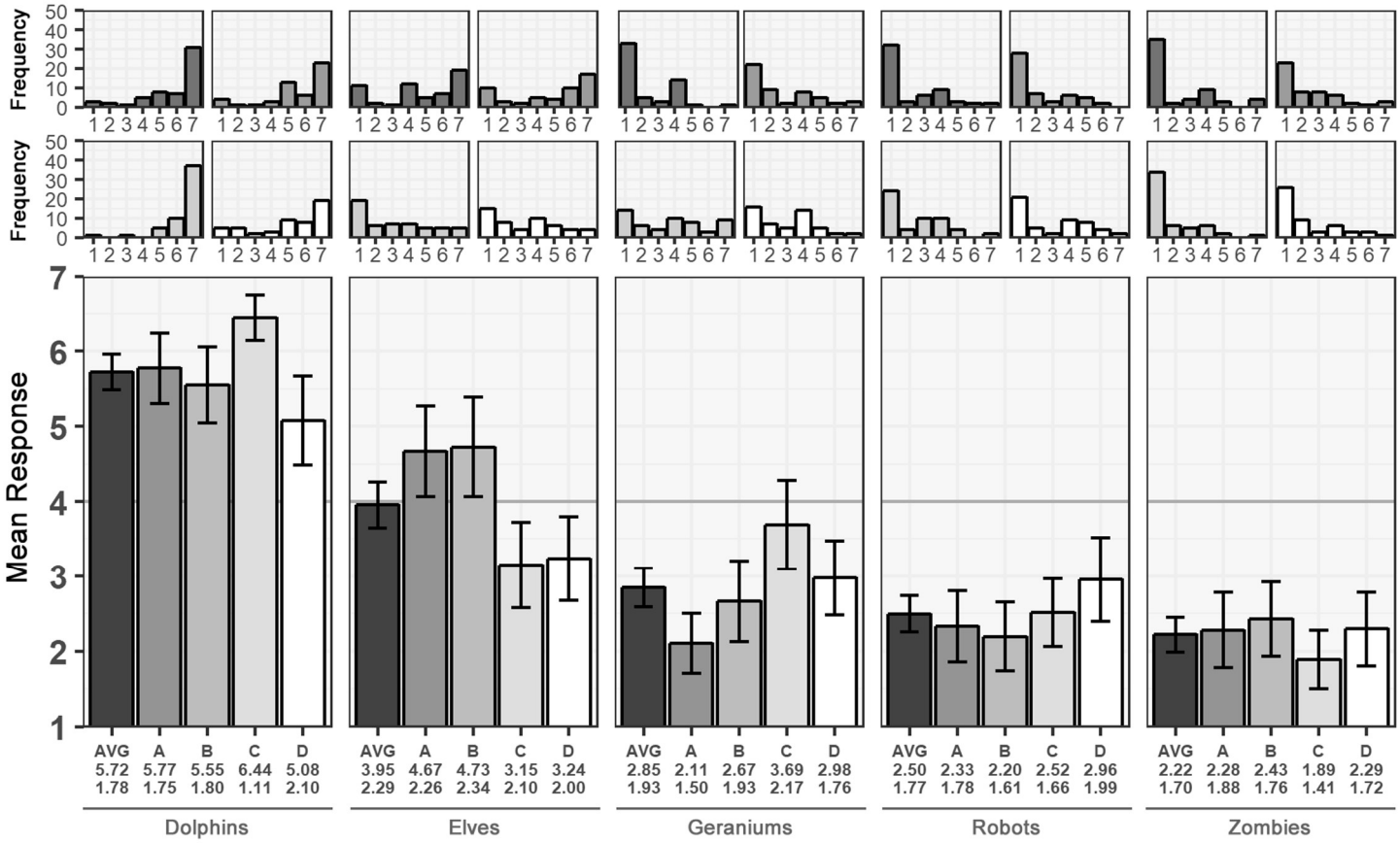

Figure C-2: Results for Study 2-B, Page 2 with means followed by standard deviations below the bar graphs; bar graphs showing 95% confidence intervals. Histograms above each bar graph show the frequency distributions of responses across participants for each condition.

Finally, our participants placed zombies on a very low rung of the ladder of consciousness visualized by Figure C-2, where zombies perch uneasily below robots and potted plants. Ratings for consciousness attributions to zombies were lower than for attributions to robots, for three of the four formulations (A, C, D), significantly so for two (C, D); there was no significant

difference between the ratings for the fourth (B).[19] Similarly, ratings for zombies were lower than for geraniums for three of the four formulations (B, C, D), significantly so for two (C, D); there was no significant difference between the ratings for the fourth (A).[20]

Findings from this study suggest that outright lack of conscious experience is inferred not so much from (clusters of) typical features of zombies as from the noun itself. This noun triggers stereotypical inferences to *lacks conscious experience*. The stereotype associated with the dominant sense of 'zombie' includes the feature *lacks conscious experience* – even though this feature had not been produced in preparatory listing tasks (see Appendix B). Association with a negative feature (like *lacks conscious experience*) can be implemented through excitatory connections to a representation of this negative feature or inhibitory connections to its positive counterpart (*has conscious experience*) – or both. Excitatory connections predict typicality ratings above neutral for attributions of the negative feature, and inhibitory connections predict typicality ratings below neutral for attributions of the positive feature.[21] Studies 3A and 3B examine these predictions.

Appendix D – Typicality of possession of conscious experience (Study 3A)

Study 3A employed a typicality rating task (i) to examine whether the positive feature *possession of conscious experience* is atypical for zombies and (ii) to confirm whether clusters of individually typical features (like T1-T3) are also collectively typical of zombies.

Participants: 148 participants were recruited as before and met the basic restrictions, with 34.5% failing one or both of the attention checks. This left 97 participants.[22]

Methods: The study included two pages with critical items, with the pages counterbalanced for order.[23] On the first page, participants rated the previously used clusters of typical and atypical zombie features (T1-T3 and A1-A3) on the same scale as in Study 2A. Items appeared in random order and included an attention check. On the second page, participants rated how typical ten additional properties are for zombies, using the same scale. Items included the four attributions of consciousness to zombies used above (A-D). To contextualise their ratings, we further included three previously studied individual typical and atypical zombie features (*move stiffly*, *infected*, *feel hungry*, and *talk*, *alive*, *feel thirsty*, respectively). Items again appeared in random order and included an attention check.

Results and discussion: Findings are presented in Figure D. Starting with the first page, ratings for T1-T3 were all significantly above mid-point.[24] In contrast, ratings for A1-A3 were

---

[19] [A] $t(56)=.17$, $p=.43$, $d=.023$ ($V=182.5$, $p=.44$, $r=.045$); [B] $t(50)=1.02$, $p=.31$, $d=.14$, two-tailed ($V=192.5$, $p=.41$, $r=.048$); [C] $t(53)=2.51$, $p=.0076$, $d=.34$ ($V=85$, $p=.011$, $r=.15$); [D] $t(50)=1.99$, $p=.026$, $d=.28$ ($V=132.5$, $p=.020$, $r=.14$).

[20] [A] $t(56)=.64$, $p=.52$, $d=.085$, two-tailed ($V=196.5$, $p=.60$, $r=.031$); [B] $t(50)=.60$, $p=.28$, d=.084 ($V=279.5$, $p=.28$, $r=.063$); [C] $t(53)=5.76$, $p<.001$, $d=.78$ ($V=46$, $p<.001$, $r=.27$); [D] $t(50)=2.09$, $p=.021$, $d=.29$ ($V=181.5$, $p=.023$, $r=.13$).

[21] While we remain agnostic here about whether laypeople possess the concept of phenomenal consciousness, we assume they represent possession and lack of a *conscious experience* feature, on some understanding, and that it is entailed by all of formulations A-D above, on their intended interpretation. We further assume that inhibition of the *possession* representation and activation of the *lack* representation will influence typicality judgments on items like A-D in the same direction (if to slightly different extent, given evident differences between these items – e.g., between (B) feelings and (C) sentience).

[22] These participants were 74.2% women (1 non-binary), mean age 34.2 years (16-75 years).

[23] To check for order effects, we conducted a mixed ANOVA with *order* as a between-subjects factor and *item* as a within-subjects factor. No significant order effects were found.

[24] [T1] $t(96)=11.18$, $p<.001$, $d=1.14$ ($V=3738$, $p<.001$, $r=.72$); [T2] $t(96)=7.20$, $p<.001$, $d=.73$ ($V=2832$, $p<.001$, $r=.58$); [T3] $t(96)=4.10$, $p<.001$, $d=.42$ ($V=2445.5$, $p<.001$, $r=.38$).

all significantly below mid-point.[25] These findings confirm that these collections of individually typical zombie features are also collectively typical of zombies. Turning to the second page, typicality ratings for each of the three typical individual features were significantly above mid-point.[26] Ratings for each of the atypical individual features were significantly below mid-point.[27] The pattern observed replicated that observed for these features in Study 2A.

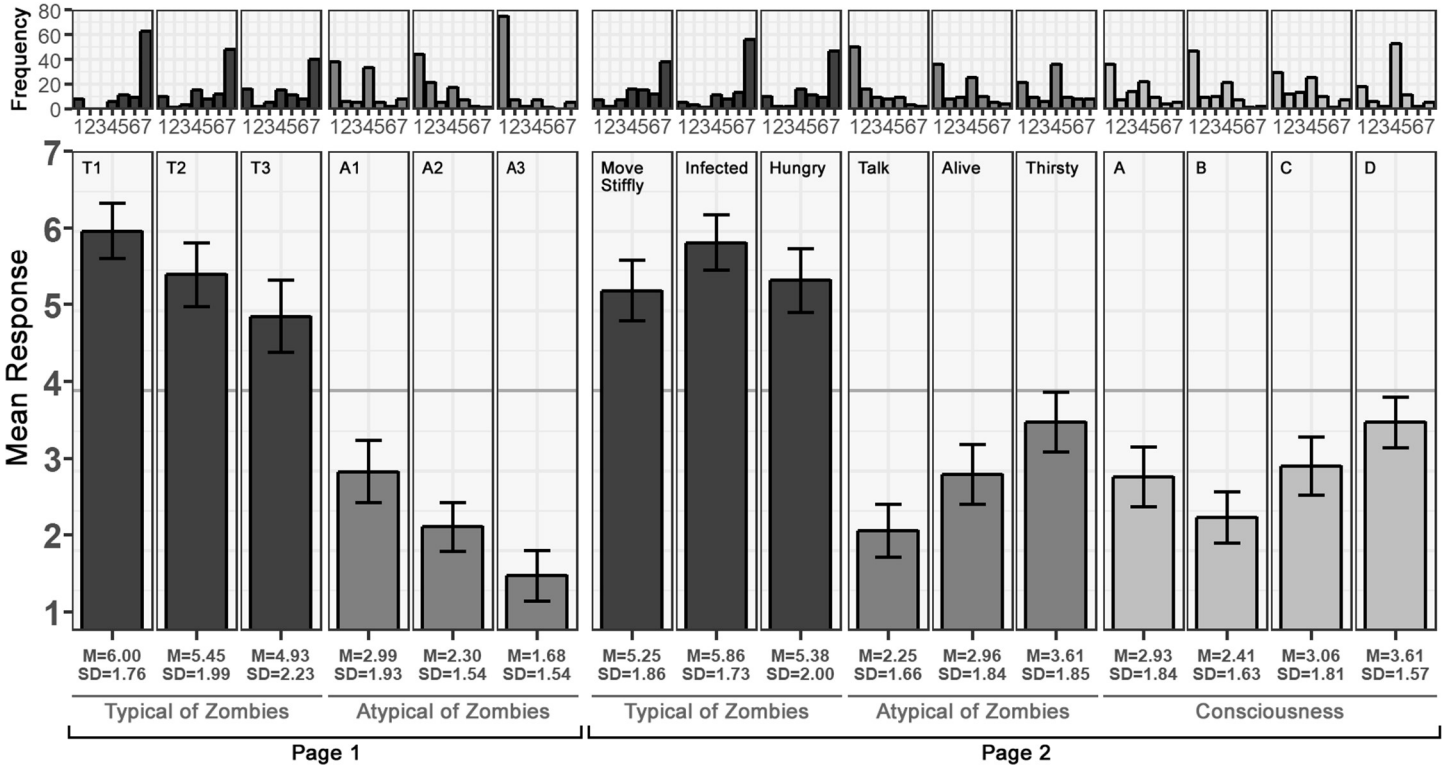

Figure D. Results for Study 3A with means followed by standard deviations below the bar graphs; bar graphs showing 95% confidence intervals. Histograms above each bar graph show the frequency distributions of responses across participants for each condition.

Turning to the crucial consciousness attributions (A-D), we examined variation between different formulations with a repeated-measures ANOVA. This showed a significant (if small) effect for *formulation* $F(3,288)=13.16$, $p<.001$, $\eta^2=.059$. Visual inspection suggests this was driven by somewhat lower ratings for B and somewhat higher ratings for D.[28] The difference is driven in part by the strikingly large proportion of participants answering '4' (neutral) for formulation D (54.6%). This is notably higher than for the related attributions A (23.7%), B (21.6%), and C (25.8%). Since each of these formulations entails D on its intended Nagelian interpretation (e.g., if a being has an 'inner mental life', there is something it is like to be it), this suggests that well over a quarter of participants in this study did not understand D as

[25] [A1] $t(96)=5.16$, $p<.001$, $d=.52$ ($V=388.5$, $p<.001$, $r=.46$); [A2] $t(96)=10.86$, $p<.001$, $d=1.10$ ($V=151.5$, $p<.001$, $d=.74$); [A3] $t(96)=14.79$, $p<.001$, $d=1.50$ ($V=254.5$, $p<.001$, $r=.81$).

[26] [move stiffly] $t(96)=6.61$, $p<.001$, $d=.67$ ($V=2768.5$, $p<.001$, $r=.56$); [infected] $t(96)=10.55$, $p<.001$, $d=1.07$ ($V=3403.5$, $p<.001$, $r=.72$); [feel hungry] $t(96)=6.80$, $p<.001$, $d=.69$ ($V=2739$, $p<.001$, $r=.57$).

[27] [talk] $t(96)=10.41$, $p<.001$, $d=1.06$ ($V=296.5$, $p<.001$, $r=.73$); [alive] $t(96)=5.57$, $p<.001$, $d=.57$ ($V=440$, $p<.001$, $r=.48$); [feel thirsty] $t(96)=2.08$, $p=.020$, $d=.21$ ($V=640$, $p=.013$, $r=.19$).

[28] This was confirmed by a series of pairwise comparisons, which showed significant differences between each pair of items except A and C $t(96)=.75$, $p=.46$, $d=.076$, two-tailed ($V=622$, $p=.54$, $r=.062$). The difference is most pronounced between B and D $t(96)=5.75$, $p<.001$, $d=.58$, two-tailed ($V=278.5$, $p<.001$, $r=.49$).

intended or at all.[29]

Crucially, however, possession of conscious experience was deemed atypical of zombies on each formulation, with all ratings significantly below mid-point: [A] $t(96)=5.74$, $p<.001$, $d=.58$ ($V=499$, $p<.001$, $r=.50$); [B] $t(96)=9.62$, $p<.001$, $d=.98$ ($V=189.5$, $p<.001$, $r=.70$); [C] $t(96)=5.11$, $p<.001$, $d=.52$ ($V=531.5$, $p<.001$, $r=.46$); [D] $t(96)=2.46$, $p=.0079$, $d=.25$ ($V=277$, $p=.0048$, $r=.17$). To illustrate, being *capable of having conscious experiences* (as per A) was deemed as atypical of zombies as *being alive* $t(96)=.14$, $p=.89$, $d=.014$, two-tailed ($V=760$, $p=.94$, $r=.0081$). We infer that the 'zombie' stereotype involves inhibitory links to the component features of conscious experience.

Appendix E – Typicality of lack of conscious experience (Study 3B)

Study 3B examined the typicality of *lack of conscious experience* more directly.

Participants: 181 participants were recruited as before and met the restrictions, with 49.7% of these failing the attention check.[30] This left 91 participants.

Methods: They used the same 7-point scale to rate how typical ten properties are for zombies. The ten items included four attributions of *lack* of consciousness:

(A*) Zombies are incapable of having conscious experiences.
(B*) Zombies lack an inner mental life, for instance, they lack feelings and emotions.
(C*) Zombies are not sentient and do not experience sensations or their surroundings.
(D*) There is nothing it feels like to be a zombie.

We added the individual typical and atypical zombie features used in Study 3A. These had been chosen so as not to prime attributions of lack of consciousness. Items appeared in random order and included an attention check.

Results: Findings are presented in Figure E. Typicality ratings for the three typical features were again significantly above mid-point.[31] Ratings for the atypical features were again significantly below mid-point.[32] Patterns for both kinds of items replicated the ones observed in the previous studies. Combining ratings for these six items for Studies 2A, 3A, and 3B, a two-way mixed ANOVA, with *study* as a between-subjects factor and *item* as a within-subjects factor showed neither a significant main effect for *study* $F(2,264)=2.25$, $p=.11$, $\eta^2=.003$ nor a significant interaction effect $F(10,1320)=1.21$, $p=.28$, $\eta^2=.004$. These replications support the reliability of our typicality ratings.

Turning to the crucial consciousness attributions, a repeated-measures ANOVA showed a significant (if small) effect for *formulation* $F(3,270)=3.97$, $p=.009$, $\eta^2=.019$. Visual inspection suggested that D* is the outlier here. The Nagelian formulation again prompted the highest proportion of neutral '4' ratings (34.1%), suggesting many participants again failed to understand it as intended or at all. We therefore ran a follow up ANOVA with D* removed.

---

[29] This peak at '4' was not so clearly observed in Study 2B which presented each participant with one of formulations A-D only. In the present study, juxtaposition of D with the arguably more readily intelligible formulations A-C may have made participants infer with the Maxim of Manner that D intended some different feature, which remained opaque to them, and rated the elusive feature 'neither typical nor atypical' because they felt unable to make any typicality (or other) judgment about it. This explanation suggests many participants in the main study felt they did not understand D at all.

[30] These participants were 64.8% women (1 non-binary), mean age 40.1 years (16-74 years).

[31] [move stiffly] $t(90)=5.56$, $p<.001$, $d=.58$ ($V=2226$, $p<.001$, $r=.51$); [infected] $t(90)=6.55$, $p<.001$, $d=.69$ ($V=2882.5$, $p<.001$, $r=.55$); [feel hungry] $t(90)=3.51$, $p<.001$, $d=.37$ ($V=1956$, $p<.001$, $r=.34$).

[32] [talk] $t(90)=11.41$, $p<.001$, $d=1.20$ ($V=113$, $p<.001$, $r=.76$); [alive] $t(90)=5.83$, $p<.001$, $d=.61$ ($V=462.5$, $p<.001$, $r=.52$); [feel thirsty] $t(90)=5.00$, $p<.001$, $d=.52$ ($V=333$, $p<.001$, $r=.46$).

This analysis did not show a significant effect for *formulation* $F(2,180)=2.28$, $p=.11$, $\eta^2=.009$. This allowed us to average across the items A*-C* that participants treated similarly. We thus found lack of conscious experience was deemed distinctly typical of zombies, with the mean rating significantly above the mid-point $t(90)=4.59$, $p<.001$, $d=.48$ ($V=2705.5$, $p<.001$, $r=.42$). Mean ratings for A*-C*, individually, were significantly above mid-point, and marginally so for D*.[33] To illustrate, being *incapable of having conscious experiences* (as per A*) was deemed as typical of zombies as *moving stiffly* $t(90)=1.38$, $p=.17$, $d=.15$, two-tailed ($V=573$, $p=.14$, $r=.15$). This supports the conclusion that *lack of conscious experience* is a component feature of the stereotype associated with the dominant sense of 'zombie'.

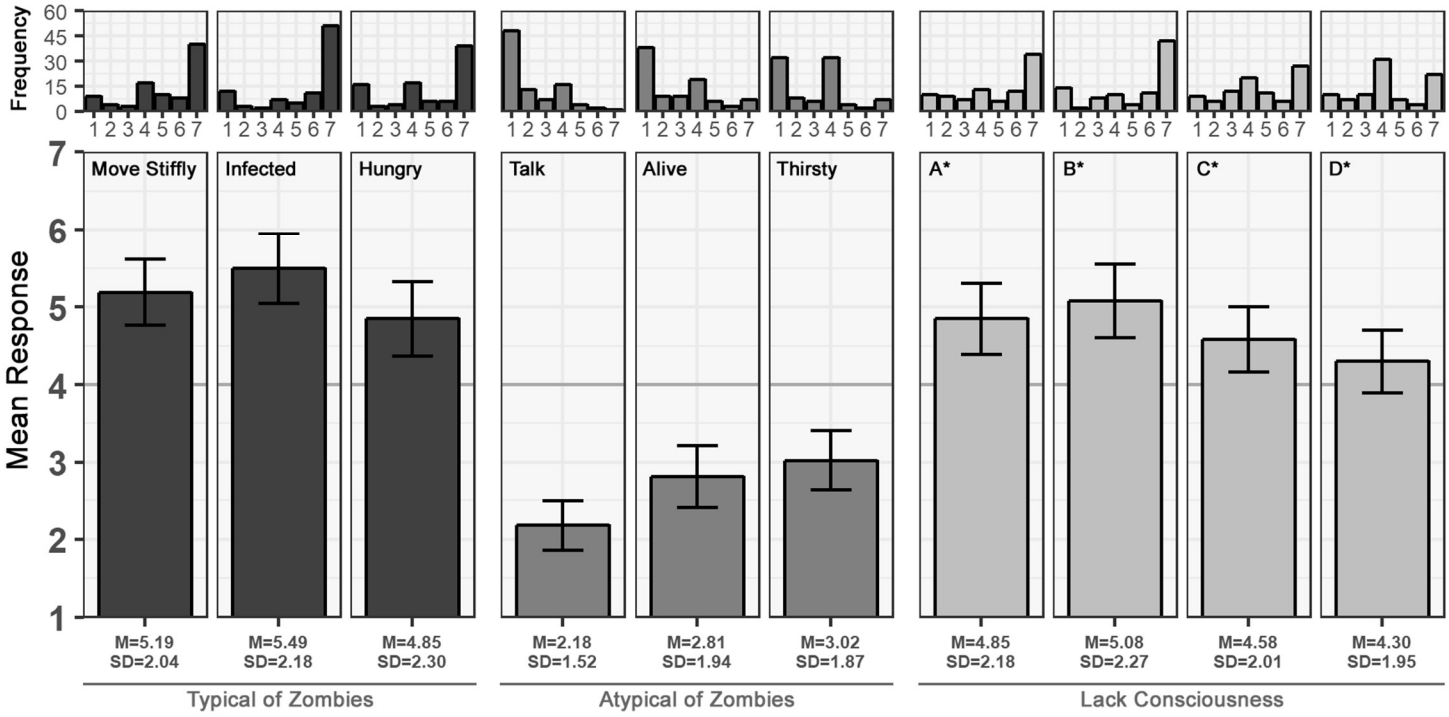

Figure E. Results for Study 3B with means followed by standard deviations below the bar graphs; bar graphs showing 95% confidence intervals, histograms shown above. Histograms above each bar graph show the frequency distributions of responses across participants.

Finally, the pattern of ratings for A*-D* inversely mirrors the pattern observed for positive attributions A-D in Study 3A (*cf.* Figures D and E). In fact, combining these items from Studies 3A and 3B, with A*-D* reverse coded, a two-way mixed ANOVA with *item* as a within-subjects factor and *study* as a between-subjects factor showed neither a significant main effect for *study* $F(1,186)=2.04$, $p=.16$, $\eta^2=.006$ nor a significant interaction $F(3,558)=.68$, $p=.56$, $\eta^2=.002$.

Appendix F – Conceivability judgments (Pre-study for Main Study)

Participants: 58 participants were recruited as before and using the same restrictions.[34]
Methods: Each participant read the following vignette describing the creation of a physical and

[33] [A*] $t(90)=3.70$, $p<.001$, $d=.39$ ($V=2251$, $p<.001$, $r=.37$); [B*] $t(90)=4.53$, $p<.001$, $d=.48$ ($V=2482$, $p<.001$, $r=.42$); [C*] $t(90)=2.76$, $p=.0035$, $d=.29$ ($V=1753.5$, $p=.0027$, $r=.27$); [D*] $t(90)=1.45$, $p=.075$, $d=.15$ ($V=1134$, $p=.050$, $r=.13$).

[34] These participants were 69.0% women, mean age 41.5 years (16-78 years).

behavioural duplicate (=**P**) that lacks conscious experience (=**~Q**):

> Here is a science-fiction story: In the future, scientists create humanoid beings. They scan the bodies of ordinary people, including their brains, at the molecular level. Using this information, the scientists then create an exact physical duplicate of a person's body, molecule by molecule. This creates a duplicate that has a body just like the original person and that behaves just like the original person. At the same time, all is dark inside for the duplicate. The duplicate lacks conscious experiences.

Participants were then asked whether they agreed or disagreed with each of two statements on a 7-point scale (anchored at 1 with 'totally disagree', at 4 with 'neither agree nor disagree', and at 7 with 'totally agree'):

> **[contradictory]** This story about duplicates is contradictory.
> **[conceivable]** It is conceivable that such duplicates might exist one day.

Results are presented in Figure F below.

The mean for [contradictory] was below the mid-point, though not significantly so $t(57)=1.47$, $p=.15$, $d=.19$, two-tailed ($V=232$, $p=.17$, $r=.20$). The mean for [conceivable] was significantly above the mid-point $t(57)=2.04$, $p=.046$, $d=.27$, two-tailed ($V=494.5$, $p=.069$, $r=.30$). The difference between the two ratings was significant $t(57)=2.43$, $p=.018$, $d=.46$, two-tailed ($V=182$, $p=.017$, $r=.31$). At first blush, this suggests that participants had a slight tendency to hold that the duplicates are not contradictory and are conceivable. However, we found only a negligible correlation between participant responses and this correlation was not significant $r=-0.040$, $p=.77$. Further, as seen in Figure D, there were a notable percentage of undecided '4' responses (39.7% for [contradictory], 34.5% for [conceivable]), including over one-quarter of participants answering '4' for both questions (25.9%). Finally, over one-quarter of participants gave problematic pairs of responses, either agreeing with both questions (17.2%) or disagreeing with both questions (8.6%).

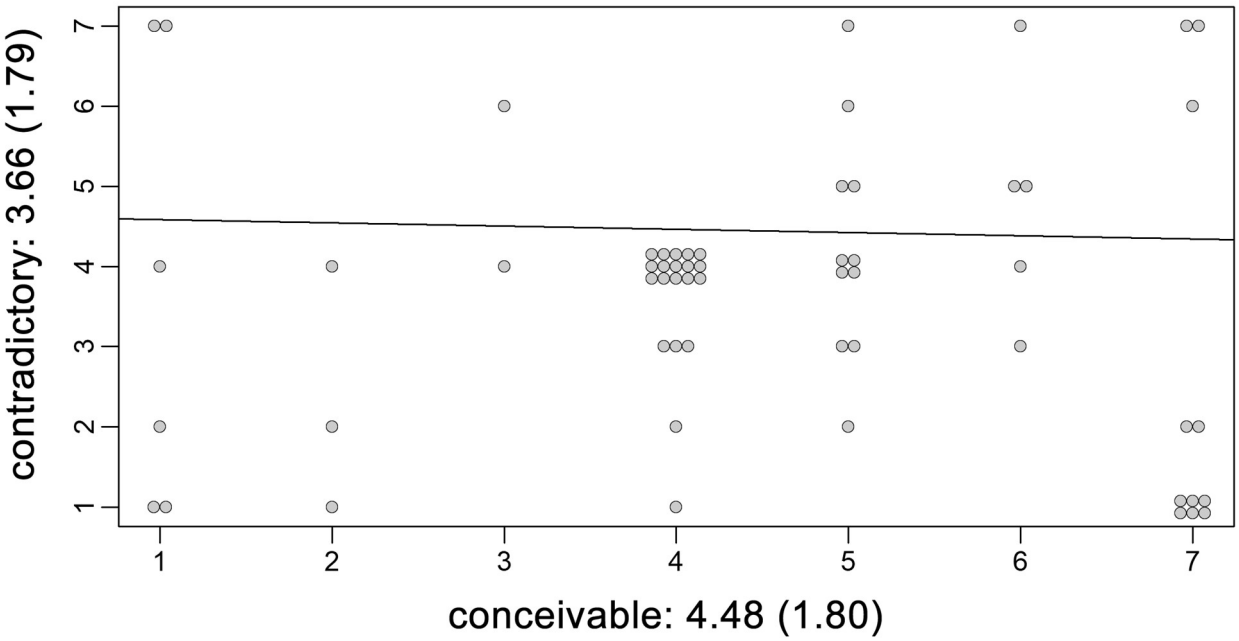

Figure F. Scatterplot for pre-study showing pattern of responses for [contradictory] and [conceivable], with linear regression line and responses spread out to show counts; mean next label with standard deviation in parentheses.

Appendix G – Effect of exclusions on effect of *term* and follow-up analysis (Main Study)

1. Effect of Exclusions

For the primary analyses for our main study that assessed H1 and H2, we excluded participants who failed the attention check on the first page ('Please select 5 for this item'), disagreed with the first comprehension on the second page ('According to the story, a zombie has a brain just like the original person's'), or agreed with the second comprehension check on the second page ('According to the story, the zombie would behave differently than the original person).

To test the impact of these exclusions, we included all participants meeting the basic restrictions (native English-speakers raised in North America, 16 years of age or older, with at most minimal training in philosophy), whether or not they passed the attention and comprehension checks. We then repeated the mixed ANOVA reported in the main text, but adding two additional between-subjects factors: *attention* (passed / failed attention check on page 1) and *comprehension* (passed / failed comprehension checks on page 2). We again found a significant main effect of *term* $F(1,630)=32.35$, $p<.001$, $\eta^2=.019$; in addition, we found significant main effects for both *attention* $F(1,630)=9.69$, $p=.0019$, $\eta^2=.006$ and *comprehension* $F(1,630)=50.90$, $p<.001$, $\eta^2=.029$. No interaction effects were found between *term* and either *attention* $F(1,630)=0.47$, $p=0.49$, $\eta^2=.000$ or *comprehension* $F(1,630)=.021$, $p=.88$, $\eta^2=.000$, and similarly there were no higher-order interactions. This means that while participants who passed or failed the checks on each page did give significantly different responses, this did not significantly change the key effect of *term* that we are interested in.

To further check the exclusions, we calculated the mean response for each condition across all items (with T1-T3 reverse coded) for each of three groups of participants: [a] participants who failed the attention check, [b] participants who passed the attention check but failed the comprehension checks, and [c] participants who passed both the attention check and the comprehension check. We found that the difference in the means between the 'duplicate' and the 'zombie' conditions increased across these three groups: [a] 0.54 (4.38 for 'duplicate', 3.84 for 'zombie'), [b] 0.58 (4.41, 3.83), [c] 0.66 (5.08, 4.41). Further, while the difference was significant for each group, the effect size was larger for [c] than for either [a] or [b]: [a] $t(159.93)=2.71$, $p=.0037$, $d=.41$; [b] $t(206.84)=3.10$, $p=.0011$, $d=0.43$; [c] $t(243.88)=4.07$, $p<.001$, $d=0.52$. A similar pattern holds looking at just the crucial consciousness attributions A-C: [a] 0.65 (4.07, 3.42), [b] 0.64 (4.30, 3.66), [c] 0.82 (5.02, 4.20). And, again, while the difference was significant for each group, the effect size was larger for [c] than for either [a] or [b]: [a] $t(169.63)=2.42$, $p=.0082$, $d=.36$; [b] $t(206.92)=2.51$, $p=.0064$, $d=.35$; [c] $t(243.41)=3.92$, $p<.001$, $d=.50$. These findings speak against the worry that the key difference between the 'zombie' and 'duplicate' conditions is due to cursory reading. In fact, quite the opposite: the effect of *term* is stronger among those paying attention and showing greater comprehension.

Finally, to test whether participants were able to conceive of philosophical zombies, we used an even stricter restriction, removing participants who answered '4' on either comprehension check. An ANOVA for the consciousness attributions A-D with *term* and *restriction* (passed/failed further restriction) as between-participant factors, and controlling for variation between participants across the items, did not show a significant main effect for *restriction* $F(1,243)=2.47$, $p=.12$, $\eta^2=.005$ or a significant interaction effect $F(1,243)=.026$, $p=.87$, $\eta^2=.000$. Similarly, excluding D there was still no main effect for *restriction* $F(1,243)=2.50$, $p=.11$, $\eta^2=.007$ and no interaction effect $F(1,243)=.008$, $p=.93$, $\eta^2=.000$. This shows that while the introduction of yet stricter restrictions to examine the conceivability question was conceptually called for (see main text, Sec.2.3), it did not notably affect ratings.

2. Follow-up analyses for H1

Given the effect of *cluster* found in the analysis in Section 2.3, we examined each item category separately. For the typical and atypical items, we ran a two-way mixed ANOVA with *term* as a between-subjects factor and *item* as a within-subjects factor. For the typical items (T1-T3), we found a main effect of *term* $F(1,245)=8.478$, $p=.0039$, $\eta^2=.017$, a main effect of *item* $F(2,490)=62.970$, $p<.001$, $\eta^2=.101$, and a marginally significant interaction $F(2,490)=2.644$, $p=.072$, $\eta^2=.004$. As predicted, follow-up tests showed that agreement is significantly higher in the zombie condition than in the duplicate condition for T1 $t(223.16)=2.92$, $p=.0019$, $d=.37$ ($W=8950.5$, $p=.0033$, $r=.17$) and T3 $t(243.9)=2.96$, $p=.0017$, $d=.38$ ($W=9232$, $p=.0018$, $r=.19$), although the difference was not significant for T2 $t(244)=.70$, $p=.24$, $d=.09$ ($W=7923.5$, $p=.29$, $r=.03$). For the atypical items (A1-A3), we again found a main effect of *term* $F(1,245)=12.29$, $p<.001$, $\eta^2=.033$ and a main effect of *item* $F(2,490)=21.634$, $p<.001$, $\eta^2=.025$. As predicted, follow-up tests showed that agreement is significantly higher in the duplicate than in the zombie condition for all three items [A1] $t(240.98)=2.40$, $p=.0086$, $d=.31$ ($W=6369$, $p=.012$, $r=.15$); [A2] $t(237.31)=3.41$, $p<.001$, $d=.43$ ($W=5783.5$, $p<.001$, $r=.21$); [A3] $t(243.44)=3.07$, $p=.0012$, $d=.39$ ($W=5955$, $p=.0013$, $r=.19$). Follow-up analyses for the consciousness items are reported in the main paper (Sect. 2.3).

Appendix H – Psycholinguistic interpretation

The main study observed small framing effects for attributions of typical and atypical zombie features (T1-T3 and A1-A3), but a medium-sized framing effect for consciousness attributions. Since conscious experience is not deemed more atypical of zombies than the other atypical features examined (A1-A3) (Study 3A, see Appendix D) and lack of consciousness no more typical than other relevant typical features (Study 3B, see Appendix E), the difference between consciousness and other attributions is unlikely to be due to different strengths of relevant stereotypical associations. Instead, it will be due to different levels of contextual support. The stereotypical inferences from 'zombie' to T1-T3 and A1-A3 clash with contextual information (physico-behavioural indistinguishability 'P') – and receive no contextual support from other parts of our vignette. Hence these inferences are largely suppressed and engender only a small difference between zombie and duplicate conditions. By contrast, our vignette, like the zombie argument, contains not only contextual information ('P') that cancels stereotypical inferences from 'zombie' to lack of conscious experience but also information that supports them ('all is dark inside'). This mitigates their suppression in the zombie condition. In the duplicate condition, where no inferences from the noun suggest lack of conscious experience, the perceived conflict between the implication (from 'P') of possession of conscious experience and the suggestion (from 'all is dark inside') of its lack is more likely to be resolved in favour of the former, by (re-) interpreting the latter (as 'full of dark thoughts and feelings'). This interpretation leaves a tension with 'P' (the average person's brain and behaviour will typically suggest less dysphoria) but renders the phrase ('all dark inside') consistent with attributions of conscious experience (whose lack is perceived as even more strongly inconsistent with ordinary physico-behavioural repertoire). This leads to larger differences between zombie and duplicate conditions, namely, the observed medium-sized framing effect for consciousness attributions.

The difference between the small framing effects observed in this study (for T1-T3 and A1-A3) and the large effects observed in previous studies of salience bias (Fischer & Engelhardt, 2019; 2020) may have two complementary explanations. All these effects result from the fact that the contextually relevant component feature of the dominant stereotype continues to pass on lateral co-activation to further component features that are frequently co-instantiated but irrelevant to the interpretation of the given subordinate use and cancelled by

contextual information (see Sec.1.3). In contrast with the previous studies, the present study employed a philosophical notion with internal tensions, where the contextually relevant component feature of the dominant zombie stereotype (*lack of conscious experience*) was simultaneously contextually cancelled (by 'P'). The relevant feature retains enough activation to notably influence ratings of consciousness attributions (resulting in the medium-sized framing effect). But its partial suppression does not leave it sufficiently strongly activated to pass on enough activation to the other features (e.g., T1-T3), for lateral cross-activation to have more than a small effect on the other judgments we elicited. If this account is correct, the fact that this study failed to observe large framing effects is due to the peculiar tension built into the philosophical notion of 'zombie' we employed.

A second explanation traces the differences between previously and currently observed effect sizes precisely to the differences between high vs low frequency and verb vs noun: Less frequent use of a word forges weaker associations between component features of the stereotype associated with it, so that less activation is passed on from contextually relevant features (like *lacks conscious experience*) to other component features (like T1-T3). This explains why framing effects observed with high-frequency words are large and the effects presently observed for attributions like T1-T3 and A1-A3 are small, and smaller than effects observed for attributions of consciousness (which are influenced directly by activation of the contextually relevant component feature). The fact that the framing effect observed for consciousness attributions was less than large could be due to the fact that stereotypes associated with nouns (other than event nouns) play a less central role than situation schemas associated with verbs in the construction of the situation models that underpin judgments about the cases described (*cf*. Melinger & Mauner, 1999; Tanenhaus & Carlson, 1989). Accordingly, noun-associated stereotypes influence these judgments to a lesser extent. If this second explanation is correct, low-frequency nouns will generally give rise to salience bias only in an attenuated form. Further research is required to decide to what extent these two potentially complementary explanations apply – and how strongly linguistic salience bias can arise from low-frequency nouns, in less peculiar and non-philosophical cases.

References

References from Appendices not included in the bibliography of the main paper.

Baroni, M., Dinu, G. & Kruszewski, G. (2014), Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 238–247.

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

Melinger, A., & Mauner, G. (1999). When are implicit agents encoded? Evidence from cross-modal priming. *Brain and Language, 68,* 185-191.

Sytsma, J. (2010). Dennett's theory of the folk theory of consciousness. *Journal of Consciousness Studies,* 17 (3-4), 107-130.

Sytsma, J. (2012). Revisiting the Valence Account. *Philosophical Topics*, 40(2), 179-198.

Sytsma, J., Bluhm, R., Willemsen, P. & Reuter, K. (2019). Causal attributions and corpus analysis. In E. Fischer & M. Curtis (Eds.), *Methodological Advances in Experimental Philosophy* (pp. 209-238). London: Bloomsbury.

Tanenhaus, M.K., & Carlson, G.N. (1989). Lexical structure and language comprehension. In W. Marslen-Wilson (Ed.), *Lexical Representation and Process* (pp. 529-561). Cambridge, MA: MIT Press.