Stereotypical Inferences:
Philosophical Relevance and Psycholinguistic Toolkit

Eugen Fischer and Paul E. Engelhardt

Abstract

Stereotypes shape inferences in philosophical thought, political discourse, and everyday life. These inferences are routinely made when thinkers engage in language comprehension or production: We make them whenever we hear, read, or formulate stories, reports, philosophical case-descriptions, or premises of arguments – on virtually any topic. These inferences are largely automatic: largely unconscious, non-intentional, and effortless. Accordingly, they shape our thought in ways we can properly understand only by complementing traditional forms of philosophical analysis with experimental methods from psycholinguistics. This paper seeks, first, to bring out the wider philosophical relevance of stereotypical inference, well beyond familiar topics like gender and race. Second, we wish to provide (experimental) philosophers with a toolkit to experimentally study these ubiquitous inferences and what intuitions they may generate. This paper explains what stereotypes are (Section 1), and why they matter to current and traditional concerns in philosophy – experimental, analytic, and applied (Section 2). It then assembles a psycholinguistic toolkit and demonstrates through two studies (Sections 3-4) how potentially questionnaire-based measures (plausibility-ratings) can be combined with process measures (reaction times and pupillometry) to garner evidence for specific stereotypical inferences and study when they 'go through' and influence our thinking.

1. Stereotypes

Many nouns (Hare et al. 2009) and verbs (Ferretti et al. 2001; Harmon-Vukic et al. 2009; McRae et al. 1997) are associated with *stereotypes*: sets of properties which come to mind first and are easiest to process, when we hear those expressions. In simple cases, we can elicit them through listing and sentence-completion tasks: Tomatoes are___. Verbs can be associated with stereotypical features of events, agents, and objects or people acted upon ('patients'). Where verbs (e.g. 'S sees X') are associated with typical features of events (S uses her eyes), agents (S has two eyes), and patients (X is in front of S), these can jointly form complex, internally structured stereotypes (a.k.a. 'generalised situation schemas') (Rumelhardt, 1978).

Stereotypical associations do not determine the extension of terms (Hampton & Passanisi 2016). But they facilitate spontaneous inferences. E.g., when you hear, 'the secretary greeted Jane', you will leap to the conclusion that Jane was greeted by a woman (Atlas & Levinson 1981). Since speakers can rely on hearers to make stereotypical inferences, they do not mention stereotypical features in talk of situations that have them (Brown & Dell 1987). But to prevent misinterpretation, we need to make deviations from stereotypes explicit, and do so in compliance with Grice's (1989) Maxim of Quantity ('Say what you must, and no more!'). At the same time, we take others to comply with Gricean maxims including Quality ('Do not

say anything you believe false or for which you lack sufficient evidence!'), and assume as a default that what they tell us is true (cp. Levine 2014); rejection as false requires effortful (rather than automatic) cognition (Gilbert 1991). Automatic stereotypical inferences therefore take us from utterances to conclusions about the world.

These inferences are (roughly) captured by the neo-Gricean *I-heuristic*: 'What is expressed simply is stereotypically exemplified' (Levinson 2000, 37). This heuristic instructs us to facilitate or find interpretations that are positive, stereotypical, and highly specific (Levinson 2000, 114):

(I-speaker) Skip mentioning stereotypical features when talking about situations which conform to the relevant stereotypes; on the other hand, make deviations from the stereotype explicit ('male secretary').

(I-hearer) In the absence of explicit indications to the contrary, assume that objects, events, agents and patients possess the features stereotypically associated with the expression.

These inferences are supported by automatic association processes in semantic memory (McRae & Jones 2013). According to the well-supported *graded salience hypothesis* (Fein et al. 2015, Giora 2003), a linguistic stimulus activates – i.e. makes more readily available for use by cognitive processes from word recognition to inferencing – *all* semantic and stereotypical features associated with the expression, in any of its uses or senses. Crucially, speed and strength of initial activation depend on the '*salience*' of the sense or use. Such salience is a function (1) of how frequently a subject is exposed to the word in this sense rather than another (inferred, e.g., from familiarity ratings or occurrence frequencies in suitable corpora) and, where appropriate, (2) of how good examples of the relevant category the word stands for in that sense (prototypicality) (Giora 2003). Features associated with the expression's most salient use are activated regardless of context. E.g., the ambiguous stimulus 'mint' activates the probe 'candy' rapidly and strongly, even where it is clearly used in a less frequent sense (prime: 'All buildings collapsed except the mint') (Simpson & Burgess 1985, Till et al. 1988). To prevent inappropriate inferences in less straightforward contexts, speakers may need to highlight deviations from stereotypes *and* from salient uses.

Largely automatic inferences in line with the (thus amended) I-heuristic result from the interplay between 'stimulus-driven' and context-insensitive activation processes and context-sensitive processes which are 'expectation-driven'. These processes initially run in parallel (Levinson, 2000). Their outputs are subsequently integrated: Processes including reinforcement and decay (Oden & Spira 1983), and more effortful suppression (Faust & Gernsbacher 1996) may mitigate initial activation, namely, in the light of contextual cues ('the secretary scratched his beard') (Sturt 2003), explicit indications of deviation from relevant stereotypes ('male secretary') (Osterhout et al. 1997), and (where appropriate) explicit marking of less salient uses ('in a special sense') (Givoni et al. 2013).

The processes reviewed occur in both language comprehension and production (Levelt 1989, Pickering & Garrod 2013, Stephens et al. 2010). They are hence set to duplicate inferences in line with the I-heuristic not only in interpersonal communication but also in the sub-vocalised speech characteristic of much philosophical thought (cp. Carruthers 2002).

2. Philosophical Relevance

Turning to the question of philosophical relevance, we will now see that the experimental study of stereotypical inferences can contribute, e.g., to (i) a key project in experimental philosophy, (ii) a traditional core concern of analytic philosophy hitherto neglected by experimental philosophers, and (iii) various projects in applied philosophy.

2.1. Experimental Philosophy

Most current experimental philosophy is practiced as cognitive science (Knobe 2016). Two distinctive strands, however, seek to engage directly with work in the mainstream of analytic philosophy: They study intuitions prompted by verbal descriptions of hypothetical cases which philosophers consider in thought experiments and arguments. The 'Concept Project' (review: Alexander 2012, pp.28-49) studies such 'case intuitions' for the purposes of conceptual analysis; the 'Warrant Project' (review: Stich & Tobia 2016) does so to assess their evidentiary value and philosophers' warrant for accepting them.

Stereotypical inferences are a likely source of many (though *not* all) such intuitions about hypothetical cases: Stereotypical inferences are bound to be made in formulating and reading verbal case-descriptions. The less informative context these descriptions provide, the more likely stimulus-driven stereotypical inferences are to go through unmodified. Such automatic inferences generate 'intuitions', in the aetiological sense from cognitive psychology (see Fischer & Collins 2015 for discussion of different notions of 'intuition' in current debates):

> *Intuitions* are judgments which are based on 'automatic inferences' (Kahneman and Frederick 2005, 268), i.e., on largely automatic cognitive processes (such as, e.g., association processes in semantic memory; Morewedge & Kahneman 2010), which duplicate inferences governed by normative or heuristic rules (e.g., the I-heuristic).

Potentially shaping many case intuitions considered by conceptual analysts, stereotypical inferences clearly matter for the Concept Project. We would now like to bring out their hitherto unexplored relevance to the Warrant Project. A first generation of 'restrictionists' studied the sensitivity of intuitions to epistemologically otiose parameters, including demographic parameters, framing, and order of presentation, and inferred lack of evidentiary value from observed sensitivities (paradigm: Weinberg et al. 2001). By contrast, the nascent 'Sources Project' (Pust 2012), aka 'cognitive epistemology' (Fischer 2014), seeks to develop psychological explanations of intuitive judgments that help us assess their evidentiary value (cp. Knobe & Nichols 2008, 8). The perhaps best-articulated approach to date develops what we have called 'GRECI explanations' (Fischer et al. 2015). Such explanations trace intuitions back to automatic cognitive processes that are **g**enerally **re**liable but predictably generate **c**ognitive **i**llusions under specific vitiating circumstances. The ultimate goal is to develop 'epistemological profiles' of cognitive processes that indicate under what circumstances we may trust the generated intuitions (Weinberg 2015). While the most prominently discussed GRECI explanations trace intuitive knowledge-attributions to a 'mind-reading competency' (e.g. Nagel 2012, Boyd & Nagel 2014, Alexander et al. 2015, Gerken & Beebe 2016), stereotypical enrichment is a domain-general language process that can shape intuitions about *any* topic.

Stereotypes have a poor reputation. But standard accounts of semantic memory (McRae & Jones 2013) suggest stereotypical inferences are generally reliable: While strength of association also depends on cognitive principles of categorical organisation (Giora 2003), associative links in semantic memory evolve in response to degree of exposure (Squire 1992, Loftus et al. 1978). They thus come to reflect co-occurrence frequencies in the subject's physical and linguistic environment. E.g., the more red tomatoes I see, or hear, read, or think about, the stronger the link between the concepts 'tomato' and 'red' becomes. If, by contrast, I am exposed to mainly (unripe) green tomatoes, the link between 'tomato' and 'green' will be strengthened, and that between 'tomato' and 'red' weakened.

This lets us identify systematic biases and specific vitiating circumstances under which stereotypical inferences are liable to generate cognitive illusions, i.e., spontaneous misjudgements that strike thinkers as compelling even after they have been corrected (Pohl 2004). For instance, a bias may arise through systematic linguistic misrepresentation and selective use of words: When people who hardly encounter academics in real life, keep reading media pieces which portray academics as absent-minded or lazy, or only cover academics who are lazy, or use the noun 'academic' only in derogatory contexts (preferring 'scientist' when reporting discoveries), these properties will come to be associated with 'academic', and tabloid readers will spontaneously infer that an otherwise unknown protagonist introduced as an 'academic' is absent-minded and lazy.

Our recent work has identified two specific vitiating conditions of particular relevance for philosophical thought experiments. Both arise from the fact that philosophers often take words with well-established uses in ordinary discourse and give them rarefied uses in which they can be applied with perfect propriety to stereotype-deviant situations. E.g., in visual contexts, appearance-verbs are ordinarily used with doxastic implications (where 'X looks F to S' implies that S is inclined to think that X is F); but philosophers of perception often use them in a 'phenomenal' sense, which serves merely to describe subjects' experience, is devoid of those implications, and can be applied to situations that tempt nobody to think that X is F (Fischer & Engelhardt 2016). Where authors fail to make the non-salient use explicit (an unwitting violation of the production-rule I-speaker), readers, and authors themselves, are liable to make stereotypical inferences licensed by the ordinary use, and do so also in inappropriate contexts, where the verb *only* applies in its special (e.g., phenomenal) use – as in the scenarios from which 'arguments from illusion' typically proceed (ibid.). Second, where differences in salience between dominant and special senses of a word are particularly pronounced, stereotypical inferences licensed by the dominant use may be made from occurrences of the special sense, even where the latter is explicitly marked (see below) – an unwitting violation of the comprehension rule (I-hearer).

In summary, stereotypical enrichment arguably is a generally reliable automatic process that generates cognitive illusions under specific vitiating conditions. The study of stereotypical inferences is therefore well placed to contribute towards GRECI explanations of intuitions philosophers have when considering verbal case-descriptions – the perhaps most ambitious strand of experimental philosophy's much-discussed Warrant Project.

2.2. Argument Analysis

At the same time, it can contribute to a key analytic task hitherto neglected by experimental philosophers, viz., the reconstruction and assessment of philosophical arguments. Brief descriptions of possible cases frequently figure as premises and trigger stereotypical inferences to conclusions presupposed in subsequent argument. The identification of contextually inappropriate inferences can expose seductive fallacies.

As an example of how such inferences may result from the above-mentioned salience differences between different senses of a word, consider the 'argument from hallucination'. This famous paradox about perception argues, against common sense, for the existence of mental objects of sense-perception ('sense-data') which separate us from the physical objects around us (Brewer 2011, Robinson 2001, Smith 2002).

> 'Let us take as an example Macbeth's visionary dagger: since we are concerned only with what is possible, the fact that this episode may be fictitious does not matter. There is an obvious [ordinary] sense in which Macbeth did not see the dagger; he did not see the dagger for the sufficient reason that there was no dagger there for him to see. There is another [viz., phenomenal] sense, however, in which it may quite properly be said that he did see a dagger; to say that he saw a dagger is quite a natural way of describing his experience. But still not a real dagger; not a physical object […] If we are to say that he saw anything, it must have been something that was accessible to him alone [viz.] a sense-datum.' (Ayer 1956/1990, 90)

The argument then generalises from such cases of visual hallucination to all cases of sight. While this generalising step has – rightly – attracted considerable criticism, the following fallacy in the first step has escaped attention: The verb 'to see' is explicitly used here in a phenomenal sense, to 'describ[e] his experience' (ibid.): 'Macbeth saw a dagger' is to mean 'Macbeth had an experience as of / like that of / seeing a dagger'. An experience (event) is being described by comparing it to that of seeing a certain physical object, and is said to be similar. This does not require that the object 'seen' be around, when the subject undergoes the experience thus described. What it does require is that the experience be in some respects similar to that of seeing a solid, physical dagger. But the case-description explicitly postulates that 'an experience of this sort is like the experience of seeing a real physical object' (ibid.), rather than that of seeing a dagger's shadow or another non-physical object. In the phenomenal sense, Macbeth can therefore be said to 'see a real dagger', but cannot be said to see a non-physical object. In the ordinary sense, he cannot be said to see anything at all. What is 'special' is the sense in which the argument uses the verb 'see' (viz. the phenomenal sense), not the object 'seen' in this sense.

We submit that the argument relies on a stereotypical inference from the most salient use of the verb 'see': In its dominant visual sense, 'S sees X' is stereotypically associated with the spatial patient property *X is in front of S* (before his eyes). Due to particularly pronounced differences in salience, an inappropriate stereotypical inference is made, despite explicit markers, from the special (phenomenal) use of the verb: Thinkers leap from

(1) ‘Macbeth sees a dagger’ to the typically tacit conclusion
(2) There is a dagger before Macbeth’s eyes.

But, by explicit assumption, there is no physical object answering the description of ‘dagger’ before Macbeth’s eyes (though perhaps elsewhere). I.e.:

(3) There is no physical dagger before Macbeth’s eyes.

Proponents of the argument infer from (2) and (3) that

(4) There is a non-physical dagger before Macbeth’s eyes.

They naturally conclude that this must be the dagger that Macbeth sees (as per 1):

(5) Macbeth sees a non-physical dagger.

Integration with intuitive introspective conceptions of the mind places this sense-datum in Macbeth’s mind, before his inner eye. On this reconstruction, the argument relies on a fallacy, namely, on a stereotypical inference from ‘see’ that is doubly inappropriate, because it (a) proceeds from a special use of the verb that does not license it and (b) is defeated by other parts of the case-description. We submit, it is, even so, automatically made and tacitly presupposed in further reasoning.

Experiments are required to find out whether competent speakers indeed make such inappropriate stereotypical inferences from ‘see’. By conducting such experiments, experimental philosophy can support otherwise controversial reconstructions of philosophical arguments – and expose hitherto unnoticed fallacies in them.

2.3. Applied Philosophy

Third, the experimental study of stereotypical inferences can contribute to the practice of experimental philosophy as applied philosophy. Such study contributes, e.g., to addressing questions of communication ethics. Ethical issues arise from the systematic bias and vitiating conditions we identified above: Once built up, stereotypical associations support automatic inferences we cannot help making. Their typically implicit conclusions are liable to be presupposed in further reasoning (Devine 1989, 2001) – and not only in philosophical argument. Any emotional or moral valence they carry, will automatically attach to the subject talked about (Rudman et al. 2001). In particular where stereotypical associations are not simultaneously shaped by direct observation of, or contact with, the phenomenon in question, sustained communicative interaction can therefore shape stereotypical associations and influence people’s beliefs, expectations, and attitudes in ways not open to their direct control. Communication strategies in advertising and especially politics are often built around achieving covert influence through build-up or reinforcement of stereotypes. This raises questions of communication ethics which have been discussed in particular for gender and ethnic stereotypes, e.g., in marketing ethics (review: Sheehan 2014, chs.4-6) and media ethics (review: Christians et al. 2016).

Some communication strategies of this kind involve the introduction of new uses of familiar lexemes, with a view to promoting certain inferences and arguments (Hallahan 2011). The experimental study of particular stereotypical inferences from specific expressions can

help us address the question of where such strategies are unethical and, indeed, judiciable – and contribute not only to communication ethics but also to applied philosophy of law.

Libel law asks us to consider how it is reasonable to interpret particular, potentially unflattering, utterances in ordinary (rather than legal) discourse:[1] Under common law, a false statement (Lunney & Oliphant 2010, 686) may be considered defamatory if its meaning is such that 'it tends so to harm the reputation of another as to lower him in the estimation of the community or to deter third persons from associating … with him' (*Restatement (Second) of Torts*, sec.559). 'The meaning of a communication is that which the recipient correctly, or mistakenly but reasonably, understands that it was intended to express' (ibid; cp. Peel & Goudcamp 2014, 362). Competent speakers/hearers employ the I-heuristic to determine this intended meaning (Levinson 2000) (see Sec.1). Stereotypical implications that are not explicitly cancelled are therefore 'correctly, or mistakenly but reasonably' taken as part of the intended meaning. Rigorously settling whether statements are defamatory can therefore require experiments to determine what automatic inferences based on unflattering stereotypical associations are made by members of the relevant community.

This question is particularly pressing – and ethically charged – where it applies to sustained communication strategies, say, of political activists. E.g., a strategy extending the use of familiar lexemes has LGBT activists use the affix '-phobic' in neologisms 'homophobic' and 'transphobic' and apply these terms to their opponents. False suggestions that someone suffers from a mental defect that would cause others not to associate with the plaintiff are among textbook examples of defamatory statements (Peel & Goudcamp 2014, 360). This raises the question whether the affix in the present labels ('-phobic') could reasonably be taken to imply that people so labelled are suffering from a mental disorder – a categorisation that is stigmatising (Angermeyer & Dietrich 2006, Parcesepe & Cabassa 2013), if to a different extent for different disorders (Crisp et al. 2000).

The question cannot be settled by looking at works first introducing these neologisms (e.g. Weinberg 1973, for 'homophobic'), since the uses and meanings of words may change when taken up by a larger community (Traugott & Dasher 2005). Nor do dictionaries settle the question: Different dictionaries offer different explanations of 'homophobic' and 'homophobia'. Some are consistent with the implication of pathology, e.g., 'irrational fear, aversion to, or discrimination against homosexuality or homosexuals' (*Merriam-Webster*, echoing Weinberg 1973). Other explanations, such as 'dislike of, or prejudice against homosexual people' (*Oxford Dictionaries Online*) are consistent with attributing only the kind of emotional and evaluative attitudes (dislike) whose even false attributions are typically not deemed defamatory (Sack 2010, 2.13).

To find out whether applying, say, the label 'homophobic' to people without a medical diagnosis is defamatory or potentially a legitimate move in political debate, further empirical investigation is required. In similar cases, forensic linguists have employed surveys to elicit

[1] The following application is therefore not touched by prominent debates about the extent to which generic processes of pragmatic enrichment are involved in courts' authoritative interpretations of legal norms (Soames 2008) or defeated by canons of legal interpretation and construction (Solum 2013), or mostly (Marmor 2014) or generally (Poggi 2011) inapplicable in legal discourse (cp. Carston 2013).

preferences for given paraphrases to complement corpus evidence (from collocations, etc.) (e.g., Durant 1996, cp. Shuy 2010). Our psycholinguistic approach allows us to make arguments more rigorous by invoking the I-heuristic and generating experimental evidence to determine what stereotypical inferences members of the relevant community make, e.g., from 'homophobic': whether they infer the agent property *is mentally ill*.

We now turn to psycholinguistic methods to examine whether, and when, specific stereotypical inferences are made from particular words, including the verb 'see' (Section 3) and 'homophobic' (Section 4) – and whether such inferences are made even in clearly inappropriate contexts, where common sense might suggest linguistic competence renders us immune to systematic missteps.

3. A Psycholinguistic Toolkit: Experiment 1

Forced-choice plausibility rankings are a useful heuristic tool to garner first evidence of such inferences. We have explained and demonstrated this convenient questionnaire-based approach in detail, elsewhere (Fischer & Engelhardt 2016). We will now explain how we can obtain more robust results by combining plausibility ratings with measures of processing effort, and will demonstrate this toolkit by presenting two studies.

Our first study explores the effectiveness of salience differences as vitiating condition of the generally reliable process of stereotypical enrichment (Section 2.1), by examining our specific hypothesis about the argument from hallucination (Section 2.2):

(H_1) Stereotypical inferences from the most salient visual use of 'S sees X' to spatial patient properties (*X is in front of S*) are made automatically, and are made even where the verb is clearly used in a different, less salient sense, devoid of spatial implications.[2]

3.1. Approach and Predictions

In plausibility-rating tasks, participants are given sentences to read or hear, and are asked to indicate how plausible they find the claims expressed. Subjective plausibility is measured through self-assessment on a 5-point Likert scale, with endpoints marked 'very implausible' (1) and 'very plausible' (5), and a 'neutral' mid-point (3), explained as 'neither plausible nor implausible, the decision feels arbitrary'.[3] When automatic inferences from an expression lead to conclusions that conflict with the rest of the sentence, this conflict reduces the subjective plausibility of the overall sentence.[4] To study stereotypical inferences (e.g., from 'S sees X' to

[2] We submit differences in salience between different senses of 'see' are due to differences in frequency and prototypicality. We used a corpus-study and a sentence-completion task, respectively, to show that in these two dimensions, visual uses outrank all other dictionary senses: They are far more salient than, e.g., epistemic and doxastic senses. Phenomenal uses are least salient. (Fischer & Engelhardt, in press)

[3] Variables measured through such ratings admit of parametric tests, when they are continuous, have a 0-point (here: 3/5 – arbitrary), and equal intervals on the scale represent equal differences in the property measured. The latter is commonly inferred from a normal distribution (Norman 2010). Where distributions are skewed, a transformation (square root, logarithm, inverse, etc.) is applied prior to parametric analysis (Tabachnick & Fidell 1989).

[4] This also holds on the experience-based approach to metacognitive judgments (Koriat 2012), which takes the subjective plausibility of a judgment to result not from reflection on its content but from features of the underlying cognitive processes: Perceived inconsistencies reduce the degree of 'fluency' or effortlessness of

X is in front of S) through plausibility-rating tasks, we can exploit this fact in two different ways, by manipulating either of two variables.

First, we can manipulate the stereotype-consistency of our sentences, and construct both *'s-inconsistent'* sentences (like 1a) where the verb is followed by a sequel that is inconsistent with the conclusion of the hypothesised inference,[5] and otherwise similar *'s-consistent'* sentences (like 2a) which are consistent with it:

1a. Jeb sees the spot on the wall behind him.
2a. Matt sees the spot on the wall facing him.

If participants make the hypothesised stereotypical inference,

[Prediction 1] participants will judge s-consistent 'see'-sentences more plausible than s-inconsistent counterparts.

However, subjective plausibility is also influenced by other factors, including inferences supported by other stereotypical associations or other selection preferences (Friederici & Frisch 2000), sentences' syntactic complexity (Gibson 1998), and the frequency of their constituent words (Trueswell 1996). S-inconsistent sentences will typically be either more complex than their s-consistent counterparts (adding negation) or use antonyms with different frequencies.

We can exclude most of these factors by, second, restricting our attention to s-inconsistent sentences and varying the main verb: For each 'see'-sentence we construct an otherwise identical sentence which replaces 'see' with a contrasting verb that lacks the stereotypical association at issue (or possesses it more weakly). To identify such a verb, we used a sentence-completion task: In a previous *norming study*, 41 undergraduate students from the University of East Anglia were asked to provide up to ten completions for stems including 'Jane sees____', 'Cathy was aware of ____', 'Bob saw____', and 'Harry is aware of____'. Both 'see' and 'aware' (='having perception or knowledge of a situation or fact', *Oxford Dictionaries Online*) can be given perceptual uses, in which certain spatial relations obtain between agents and patients, and non-perceptual uses, in which they don't ('I see your point', 'He's aware of the opportunity'). We found that whereas over 93% of relevant completions gave 'see' a perceptual use, only 46% did so for 'aware', yielding random completion-preferences for 'aware'. We inferred that 'S is aware of X' is associated more weakly (if at all) with the property *X is in front of S*. For each s-inconsistent 'see'-sentence we therefore constructed an otherwise identical 'aware'-sentence, e.g.:

1b. Jeb is aware of the spot on the wall behind him.

Each participant is presented with one member of each pair (1a or 1b), and items are rotated across participants. If they make the hypothesised inferences from 'see' (which are not, or not so well, supported by 'aware'),

---

the comprehension process (Just & Carpenter 1977), and fluency serves as a meta-cognitive cue for plausibility (Thompson et al. 2011; cp. Alter & Oppenheimer 2009).

[5] These sentences are not inconsistent *simpliciter*: In 1a, e.g., Jeb might see the spot in a mirror.

[Prediction 2] participants will judge s-inconsistent 'see'-sentences less plausible than 'aware'-sentences in the same sentence frames.

Third, we can combine the two approaches in a 2 × 2 design (s-consistent/s-inconsistent context × see/aware). This requires 'aware'-counterparts also for s-consistent 'see'-sentences:

2b. Matt is aware of the spot on the wall facing him.

If an ANOVA reveals an interaction between verb and contextual consistency, we have further evidence that both (i) the nature of the context (s-consistent vs. s-inconsistent) and (ii) properties of the verb (e.g., stereotypical associations) influence subjective plausibility. Further t-tests then allow us to establish whether differences in plausibility are significant between different conditions (e.g., s-consistent vs. s-inconsistent).

We built on this combined approach to examine our hypothesis (H_1) that spatial inferences are made also from less salient uses of 'see', where they are inappropriate. As less salient sense we chose the epistemic sense ('understand/know something', *Macmillan Dictionary*, sense 4); this is still more salient than the phenomenal sense (see Fn.2), to which any positive findings will therefore apply *a fortiori*. To follow up (H_1), we tested the more specific

[Predictions 1 and 2*]* that predictions 1 and 2, respectively, will hold true in both visual *and* epistemic contexts, where 'see' is used in a visual and a purely epistemic sense, respectively.

We used the combined approach outlined but manipulated also the concrete vs. abstract nature of the object nouns of 'see' and 'aware' (e.g., 'spot on the wall' vs. 'problems'). In the absence of contextual cues, concrete objects invite visual interpretations of 'see', while abstract objects (things that are not literally visible) invite purely epistemic interpretations. We thus added s-inconsistent and s-consistent sentences like:

3a. Jack sees the problems he left behind.[6]
3b. Jack is aware of the problems he left behind.
4a. Joe sees the problems that lie ahead.
4b. Joe is aware of the problems that lie ahead.

This yields a 2 × 2 × 2 (context × verb × object) design. Items belonging to all conditions are presented to each participant. This within-subject design can pick up medium-sized effects and establish positive results with as few as 20 participants, while approximately 50 are required to establish conclusive negative results or null-findings (Tabachnick & Fidell 1989).

The plausibility differences predicted by [1*] and [2*] provide evidence that stereotypical inferences are made. Where such inferences are made in inappropriate (e.g. epistemic or s-inconsistent) contexts, their conclusions need to be suppressed or integrated into a consistent situation model, leading to increased response times (Zwaan 1999, Zwaan & Radvansky

[6] In these items, also the spatial expressions are used metaphorically (e.g. 'left behind' = 'in the past'). However, if spatial inferences are made, the impression of a conflict will arise from the fact that salient spatial/literal meanings are immediately activated and retained for metaphor interpretation (Giora & Fein 1999, Fein et al. 2015, cp. Gentner et al. 2002).

1998). Greater response times can therefore provide further evidence of inferences, and we measured reaction times as well.

To show that those inferences were made as participants heard the sentence, rather than during the plausibility assessment, we used pupillometry (Kahneman 1973, Laeng et al. 2012): When we expend cognitive effort, e.g., to overcome comprehension difficulties, our pupils reliably increase in diameter; pupil dilation is an index of effort – such as that resulting from a clash between a stereotypical inference from a word and the textual sequel. The pupil takes approximately 1 second to expand to its maximum size (Engelhardt et al. 2010). We therefore measured pupil dilation during a 1 second *'offset time window'* after the sentence had been heard, and before the rating task was set. To obtain further evidence that contextually inappropriate stereotypical inferences are made during initial sentence comprehension, we compare dilations for s-inconsistent 'see'- and 'aware'-sentences. Since these sentences only differ in their main verb, it is reasonable to trace differences in pupil diameter to differences in stereotypical associations of, and inferences from, those verbs. Our hypothesis about stereotypical inferences from 'see' predicts critical dilations for s-inconsistent 'see'-, but not 'aware'-sentences.

3.2. Methods

21 undergraduate students from the University of East Anglia (UEA) participated for course credit. All were native speakers of English. Each participant heard 76 sentences. These included 24 critical sentences, namely, three in each of the eight conditions:

See, visual, s-inconsistent (e.g., 1a)
See, visual, s-consistent (2a)
See, epistemic, s-inconsistent (3a)
See, epistemic, s-consistent (4a)
Aware, visual, s-inconsistent (1b)
Aware, visual, s-consistent (2b)
Aware, epistemic, s-inconsistent (3b)
Aware, epistemic, s-consistent (4b)

Participants were presented with a fixation cross on a computer screen while hearing each sentence. Pupil dilation was measured with an Eyelink 1000, during a 1000ms time window after the end of each sentence. After this, the plausibility rating scale appeared and participants rated the sentences on a scale from 1 to 5, by pressing the corresponding key on the keyboard. Response times were measured from the appearance of the rating scale. Since we were interested in intuitive judgments, we asked participants to respond as quickly as possible, to secure responses in less than 5 seconds, before controlled processes may modify automatic cognition (DeNeys 2006).

3.3. Results

Results confirmed our predictions, including 1* and 2* (see Figure 1).

(Figure 1, p.25)

Plausibility: One button press error was made, and removed from the dataset. Due to skewed distributions, the mean plausibility ratings were transformed using the square root transform. A 2 × 2 × 2 (object × context × verb) repeated measures ANOVA showed a significant interaction ($F(1,20)=51.25$, $p<.001$, $\eta^2=.72$). In order to decompose this interaction and locate the significant differences, we considered the s-consistent and s-inconsistent sentences separately (see Figure 1). Where the 2 × 2 (object × verb) interaction was significant, we followed up with four paired-samples t-tests.[7] There were significant interactions for both s-consistent ($F(1,20)=4.83$, $p=.04$, $\eta^2=.19$) and s-inconsistent sentences ($F(1,20)=43.34$, $p<.001$, $\eta^2=.68$). For the s-consistent contexts, only the difference between visual and epistemic 'see'-items was significant ($t(20)=3.25$, $p=.004$, $\eta^2=.35$) and visual 'see'-sentences had higher ratings. For the s-inconsistent contexts, three of four paired comparisons were significant (visual-aware vs. visual-see $t(20)=11.25$, $p<.001$, $\eta^2=.86$; epistemic-aware vs. epistemic-see $t(20)=2.91$, $p=.009$, $\eta^2=.30$; and visual-see vs. epistemic-see $t(20)=-8.15$, $p<.001$, $\eta^2=.77$); one was marginally significant (visual-aware vs. epistemic-aware $t(20)=-2.25$, $p=.036$, $\eta^2=.20$). As per Prediction 2*, s-inconsistent 'see'-sentences with visual and epistemic objects were thus both deemed significantly less plausible than their 'aware'-counterparts. To follow up Prediction 1*, we also ran t-tests across s-consistent and s-inconsistent conditions. As predicted, s-inconsistent 'see'-sentences were deemed significantly less plausible than their s-consistent counterparts, both when used with visual objects ($t(20)=14.15$, $p<.001$, $\eta^2=.91$) and when having epistemic objects ($t(20)=3.47$, $p=.002$, $\eta^2=.38$).

Reaction Time: A 2 × 2 × 2 (object × context × verb) repeated measures ANOVA showed a marginal interaction ($F(1,20)=3.33$, $p=.08$, $\eta^2=.14$). As before, to decompose the (marginal) interaction, we considered s-consistent and s-inconsistent items separately. A 2 × 2 (object × verb) repeated measures ANOVA showed a marginal interaction for s-consistent items ($F(1,20)=3.90$, $p=.06$, $\eta^2=.16$). Participants took significantly longer to respond to s-consistent 'see'-sentences with epistemic, rather than visual objects ($t(20)=-4.43$, $p<.001$, $\eta^2=.50$). Perhaps surprisingly, they also took marginally longer to rate s-consistent 'aware'-sentences with epistemic, rather than visual objects ($t(20)=-2.16$, $p=.043$, $\eta^2=.19$). In addition, response times were marginally longer for s-consistent 'see'-sentences with epistemic objects than for their 'aware'-counterparts ($t(20)=-2.04$, $p=.055$, $\eta^2=.17$). The s-inconsistent items did not show a significant nor marginal interaction ($p=.36$), and no further t-tests were run. However, for all four kinds of s-inconsistent sentences (see-visual, see-epistemic, aware-visual, aware-epistemic), mean reaction times were numerically higher than for their s-consistent counterparts,[8] and numerically almost identical for s-inconsistent 'see'-sentences with visual and epistemic objects (Figure 1).

---

[7] At a referee's suggestion, we corrected for multiple comparisons, and thus use 0.0125 as the significance level for paired comparisons in this study. While this helps to guard against inflated Type I error, there is a robust debate about whether this is too conservative (e.g., Armstrong, 2014; Cabin & Mitchell, 2000; Nakagawa, 2004). We may therefore err on the side of caution in referring to *p*-values between .0125 and .05 (rather than the customary .05 and .08) as marginally significant, for paired comparisons.

[8] This finding was confirmed by a robust main effect of context on reaction times ($F(1,20)=26.00$, $p<.001$, $\eta^2=.57$).

Pupillometry: A 2×2×2 (object × context × verb) repeated measures ANOVA showed that the interaction was not significant ($F(1,20)=1.61$, $p=.22$, $\eta^2=.08$). Since the numeric mean values (Figure 1) display the same pattern as the reaction times, and in all conditions numerical plausibility-differences between 'see'- and 'aware'-items are mirrored by reverse differences in both reaction times and pupil diameters, we put this lack of significance down to the small sample size and low number of items per condition in this pilot study, and still regard it as illustrating a useful combination of plausibility ratings with psycholinguistic processing measures. In an exploratory spirit, we therefore ran follow-up analyses in which we considered whether the pupil diameter was significantly different in the 1 second time window following sentence offset as compared to pupil size during sentence reading. To do so, we ran one-sample t-tests with a test value of 1. This value would indicate that the mean pupil diameter was the same during and after reading the sentence. As predicted, participants' mean pupil size increased after reading s-inconsistent 'see'-sentences with visual *and* epistemic objects, to a significant and marginally significant extent, respectively ($t(20)=2.22$, $p=.038$ and $t(20)=1.97$, $p=.063$, dispensing with correction for multiple comparisons, in this exploratory setting). As further predicted, analogous 'aware'-sentences with visual and epistemic objects did not prompt even marginally significant increases, ($t(20)=-.019$, $p=.985$ and $t(20)=.932$, $p=.362$, respectively).

3.4. Discussion

We will now discuss how these results support our hypothesis and its philosophical application, and use this discussion to illustrate how processing measures help interpret plausibility data. We hypothesised that, during reading or listening, competent speakers make stereotypical inferences from visual *and* less salient (e.g., epistemic or phenomenal) uses of 'see' to spatial conclusions inappropriate for those less salient uses. In explaining the otherwise puzzling reasoning in the argument from hallucination, we further assumed that, in such inappropriate contexts, these conclusions are implicitly presupposed in further reasoning. This means that, despite their glaring contextual impropriety, they are only partially suppressed and remain available for further inferences.

The predicted differences in plausibility ratings (as per Predictions 1* and 2*) provide evidence that the hypothesized stereotypical inferences are made from both visual and less salient epistemic uses of 'see'. Spatial inferences may be contextually inappropriate either because they are made from an epistemic use of 'see' or because, though proceeding from a visual use, they are explicitly cancelled by the s-inconsistent context. In either case, conclusions from inappropriate inferences need to be suppressed. This increases reaction times. Evidence of such extra effort, and thus (indirectly) of contextually inappropriate initial inferences, is therefore provided by the observation that in all other conditions reaction times for 'see'-sentences are higher than in the visual s-consistent condition, in which spatial inferences do not require suppression. Indeed, mean response times are almost identical for s-inconsistent 'see'-sentences with visual and epistemic objects, where they are numerically higher than in all other conditions. The same holds true of pupil dilations, which are (marginally) significant for s-inconsistent 'see'-sentences with either kind of object. This strongly suggests that directional inferences in need of subsequent suppression are made from

epistemic uses of ‘see’ no less than from visual uses of the verb. Significant increases in pupil dilation provide evidence that these inferences are initially made when hearing the relevant sentences, rather than when assessing their plausibility.

To assess our further assumption that inappropriate conclusions are merely partially suppressed, we consider responses to ‘aware’-sentences across all conditions, and compare them to the responses to ‘see’-sentences (see Figure 1). Based on our previous norming study (above), we had tentatively assumed that ‘aware’ would be largely object neutral and spatially neutral, i.e., that responses to ‘aware’-sentences would be hardly affected by manipulations of object (visual vs. epistemic) and spatial stereotype-consistency (s-consistent vs. s-inconsistent). Even so, we found that s-consistent ‘aware’-sentences with epistemic objects prompted significantly longer reaction times than those with visual objects, and that ‘aware’-sentences with visual objects prompted marginally longer reaction times when followed by an s-inconsistent sequel than by an s-consistent sequel ($t(20)=-2.55$, $p=.019$, $\eta^2=.25$). We suggest that participants assume as a default that subjects of awareness will become aware of X by seeing X, so that ‘aware’ is associated – like ‘see’, but more weakly – with a spatial situation schema; this supports inferences which subsequently need to be suppressed (i) in purely epistemic contexts (with objects like ‘problems’ that do not literally stand in spatial relations) and (ii) in s-inconsistent contexts, resulting in the observed higher reaction times.

On this basis, evidence of the extent (partial vs. complete) to which suppression succeeds is provided by plausibility ratings: s-consistent ‘aware’-sentences were not deemed significantly less plausible when having epistemic objects than with visual objects ($t(20)=1.76$, $p=.094$, $\eta^2=.13$), despite the observed longer reaction times. This suggests initially activated but contextually inappropriate elements of the spatial situation schema were completely suppressed in these epistemic contexts, and no longer interfered with plausibility judgments. By contrast, ‘aware’-sentences with visual objects were deemed significantly less plausible when they had an s-inconsistent sequel rather than an s-consistent sequel ($t(20)=5.39$, $p<.001$, $\eta^2=.59$). This suggests merely partial suppression of directional inferences from ‘aware’ in contexts with visual objects.

The comparison with responses to ‘see’-sentences is instructive: As noted, reaction times for ‘see’-sentences (as for ‘aware’-sentences) are higher in all other conditions than in the visual s-consistent condition, in which spatial inferences do not require suppression. But now we also observe significantly lower plausibility ratings, in all other conditions. The drop is slight between s-consistent ‘see’-sentences with visual and epistemic objects, pronounced between the latter and their s-inconsistent counterparts, and dramatic between s-consistent and s-inconsistent ‘see’-sentences with visual objects (Figure 1). This suggests increasingly partial suppression of inappropriate stereotypical inferences from ‘see’ in the epistemic s-consistent, the epistemic s-inconsistent, and the visual s-inconsistent conditions. The low mean plausibility rating for ‘see’-sentences in the latter condition (1.63) indicates that where ‘see’ goes with a visible object, many participants did not suppress the spatial conclusions at all and rejected the stereotype-inconsistent sentences as outright inconsistent (rating 1; 36 times out of 62 trials). Mean plausibility ratings were significantly above mid-point for s-inconsistent ‘see’-sentences with epistemic objects (Mean = 3.54, $t(20)=2.36$, $p=.028$). This suggests a higher level of success in suppressing inappropriate inferences in this condition.

Simultaneously, the significant drop in plausibility between s-consistent and s-inconsistent 'see'-sentences with epistemic objects, despite numerically higher reaction times, suggests merely partial suppression specifically of directional inferences, in this key condition.

This conclusion is reinforced by comparison with 'aware': s-inconsistent 'aware'-sentences with epistemic objects were not deemed significantly less plausible than their s-consistent counterparts ($t(20)=1.83$, $p=.082$, $\eta^2=.14$), despite marginally higher reaction times ($t(20)=-2.34$, $p=.03$, $\eta^2=.22$). This is indicative of complete suppression of directional conclusions. But s-inconsistent 'see'-sentences with epistemic objects were judged significantly less plausible than their 'aware'-counterparts. We infer that directional conclusions from 'see' are not as completely suppressed as similar conclusions from 'aware' (where underpinning stereotypical associations are weaker).

This study provided evidence that competent speakers (i) make spatial inferences licensed by the highly salient visual sense of 'see', also from less salient uses, where they are inappropriate, and (ii) fail to suppress the conclusions obtained more than partially, so that these conclusions are liable to be presupposed in further reasoning (Devine 1989, 2001). As explained, the positive findings obtained with the epistemic use of 'see' are bound to apply to all other uses with similar and lesser salience (e.g., doxastic and phenomenal uses, see Fn.2). Our first study thus contributes, quite generally, to identifying pronounced differences in salience between different uses of a word as a vitiating condition under which the generally reliable process of stereotypical enrichment predictably misfires (Section 2.1). More specifically, it supports our reconstruction of the 'argument from hallucination', which exposes a hitherto unrecognised fallacy at the root of this classical paradox (Section 2.2). Our second study addresses our topic from the applied philosophy of law (Section 2.3).

4. A Necessary Combination: Experiment 2

Online measures like pupillometry, reading-time measurements (Klin et al. 1999, Harmon-Vukic et al. 2009) with eye-tracking (Patson & Warren 2010), and electrophysiological measurements of event-related brain potentials (Kutas & Federmeier 2011), are frequently regarded as psycholinguistic gold-standard, and offline measures (like plausibility ratings) treated as second best. This paper argues for a combination of both: Experiment 1 showed that adding online to offline measures helps determine whether stereotypical inferences are not merely made in language comprehension but potentially affect further thought. Experiment 2 will show we need to add (supposedly 'second best') offline to online measures, to tackle this question. The study uses our toolkit to investigate stereotypical inferences from 'S is homophobic' to the potentially libellous conclusion that S is mentally ill.

4.1. Methods

51 UEA undergraduate students participated for course credit. All were native speakers of English. Each heard 96 items. These included 18 critical items (3 per condition) from a list of 54 rotated across subjects. In critical items, attributions of homophobia were followed by attributions of either a mental disorder (*'disorder-consistent'*) or mental health (*'disorder-inconsistent'*). To control for level of social propriety, we used three different formulations for each sequel:

disorder-consistent

1. John is homophobic. He has a mental health condition.
2. Tom is homophobic. He has mental health issues.
3. Tim is homophobic. He is mentally ill.

disorder-inconsistent

4. Jack is homophobic. He has no mental health condition.
5. Jim is homophobic. He has no mental health issues.
6. Joe is homophobic. He is mentally healthy.

To control for gender stereotypes, each participant heard, for each item with a male subject, one otherwise identical item with a female subject ('Jane… She…'), and one with a gender-neutral subject. The latter combined two unisex names[9] with the pronoun 'they' ('Charlie and Brett are homophobic. They are mentally ill.').

In text-comprehension, discourse context may activate schemas that are not associated with any particular words in the given sentence but organise beliefs about the kind of situation or phenomenon under discussion (Metusalem et al. 2012). Inferences from attributions of homophobia may therefore be influenced not only by stereotypical associations with the word, but also by implicit theories that might attribute adverse behaviours and attitudes towards homosexuals to some mental disorder. Indeed, we expected some of our undergraduate participants to implicitly hold this *'disorder theory'*. We therefore manipulated also the initial verb phrase: Further critical items attributed strong dislike or prejudice. E.g.:

1' John strongly dislikes homosexuality. He has a mental health condition.
1* John has strong prejudices against homosexuals. He has a mental health condition.

This yields a 2 (disorder-consistent / disorder-inconsistent context) × 3 (verb: dislikes / is homophobic / has prejudices) design. We elicited plausibility ratings and measured reaction times and pupil dilation as in Experiment 1, but without requesting speedy responses (since our interest now is not restricted to intuitions).

The design indicated allows us to identify implicit 'disorder theorists' as those who judge disorder-consistent items more plausible than disorder-inconsistent items, even when the word 'homophobic' is not used (as in 1' and 1*), i.e., in the dislike- and prejudice-conditions. By contrast, *'innocent participants'* who do not regard dislike or prejudice against homosexuals as indicative of mental disorder should judge disorder-consistent items in these conditions (like 1' and 1*) less plausible than their disorder-inconsistent counterparts ('…has no mental health condition'), simply because the baseline probability that someone has mental health issues is far lower than that someone is mentally healthy.[10] We therefore use higher ratings for disorder-inconsistent items across dislike- and prejudice-conditions to identify the 'innocents'.

---

[9] From: http://www.babycentre.co.uk/l25008043/top-30-unisex-names-photos (7.9.2016)

[10] Media coverage that might inform participants' assessments suggests that ca. 25% of the UK population experience mental health issues in any given year. E.g.: http://www.wired.co.uk/article/mental-health-stats-uk (last accessed 24.10.2016).

Libel charges require the hypothesis that

(H_2) Stereotypical inferences are made from 'S is homophobic' to maintained attributions of mental disorder or illness, to S.

This 'prosecution hypothesis' predicts:

[Prediction A] Even innocent participants will rate disorder-consistent 'homophobic'-items (like 1-3) more plausible than disorder-inconsistent counterparts (like 4-6).

[Prediction B] At least innocent participants will judge disorder-inconsistent 'homophobic'-items (like 1) less plausible than 'dislike' and 'prejudice' sentences (like 1' and 1*) in the same sentence frames.

[Prediction C] Innocent participants display significant pupil dilations after disorder-inconsistent items with 'homophobic', but not with 'dislike' or 'prejudice', nor for disorder-consistent 'homophobic'-items.

4.2. Results

Results refuted prediction A, failed to support B, and confirmed C (see Figure 2) – suggesting the hypothesised inferences are initially made, but then 'drowned out' by background beliefs.

(Figure 2, p.26)

Plausibility: Using the above criteria, we identified 17 implicit disorder theorists, 21 innocent participants, and 13 participants who deemed disorder-consistent dislike- and prejudice-items as equally plausible as their disorder-inconsistent counterparts. Indeed, participants in this 'no variance' group tended to give the same rating to all critical items, across all six conditions, with values ranging from 1 (for all) to 5 (for all). We concluded that these participants adopted a response strategy, rather than engaging with the task,[11] and disregard their responses henceforth.

For the *disorder-theory group*, a 2 × 3 (context × verb) repeated measures ANOVA revealed no significant interaction ($F(2,32)=.75$, $p=.483$, $\eta^2=.05$), but a main effect of context ($F(1,16)=16.60$, $p=.001$, $\eta^2=.51$). Following this up with paired-samples t-tests, we found that disorder theorists deemed disorder-consistent items more plausible than disorder-inconsistent items, across all three verb conditions (dislike: $t(16)=4.66$, $p=.000$; prejudice: $t(16)=3.41$, $p=.004$; homophobic: $t(16)=2.99$, $p=.009$). The differences in the 'dislike' and 'prejudice' conditions are artefacts of the group definition; the expected difference in the 'homphobic' condition provides a welcome sanity check.

Also for the key *innocent group*, we found no significant interaction ($F(2,40)=.87$, $p=.427$, $\eta^2=.04$), but a main effect of context ($F(1,20)=22.18$, $p=.000$, $\eta^2=.53$). Follow-up t-tests confirmed that – in line with the group definition – innocent participants deemed disorder-inconsistent items with 'dislike' and 'prejudice' more plausible than their disorder-consistent counterparts (dislike: $t(20)=-3.84$, $p=.001$; prejudice: $t(20)=-4.23$, $p=.001$). Crucially,

[11] Markedly lower reaction times and lack of significant pupil dilations for these participants further support this conclusion.

however, innocent participants also found ‘homophobic’ items more plausible when they were disorder-inconsistent rather than disorder-consistent ($t(20)=-2.16$, $p=.043$). Since correction for multiple comparisons (see Fn.7) now sets the critical significance level at $p=.0167$, this numerical difference is not significant. But it suffices to refute Prediction A, which predicts a significant difference in the opposite direction. To support Prediction B, we would need to make comparisons between ratings of items using different verbs. In the absence of a main effect of verb ($F(2,40)=.96$, $p=.393$, $\eta^2=.05$), this is illegitimate, and the prediction remains unsupported. (Exploratory paired-samples *t*-tests reveal – against the prediction – no significant difference between the ratings of disorder-inconsistent ‘homophobic’-items and their ‘dislike’ and ‘prejudice’ counterparts.)

Reaction Time: A 2 × 3 (context × verb) repeated measures ANOVA showed that the interaction was not significant for either disorder theorists ($F(2,32)=2.57$, $p=.09$, $\eta^2=.14$) or innocent participants ($F(2,40)=2.40$, $p=.104$, $\eta^2=.11$). There was no main effect of verb or context, for either group (all *p’s* >.29). This prevented comparisons between conditions. Note, however, that mean reaction times across conditions were markedly longer for both disorder theorists (2100ms) and innocent participants (2273ms) than for participants in the speeded task of Experiment 1 (where mean reaction times remained below 1500 ms in all conditions).

Pupillometry: For the *disorder theory group*, a 2 × 3 (context × verb) repeated measures ANOVA revealed a significant interaction between verb and context ($F(1,16)=8.23$, $p=.011$, $\eta^2=.34$), due to differential performance in the disorder-inconsistent conditions ($F(1,16)=11.23$, $p=.004$, $\eta^2=.41$), and a marginal main effect of verb ($F(1,16)=4.14$, $p=.059$, $\eta^2=.21$). The only items to cause significant pupil dilations in the 1-second offset period were ‘homophobic’-items with disorder-inconsistent sequels ($t(16)=3.19$, $p=.006$). Mean maximum pupil size in this condition was significantly and marginally larger than for disorder-inconsistent items with ‘dislike’ ($t(16)=-3.57$, $p=.003$) or ‘prejudice’ ($t(16)=2.20$, $p=.043$), respectively.

For the *innocent group*, we found a marginal interaction between verb and context ($F(1,20)=4.17$, $p=.055$, $\eta^2=.32$), due to differential performance in the disorder-inconsistent contexts ($F(1,20)=12.32$, $p=.002$, $\eta^2=.38$), and a main effect of verb ($F(1,20)=9.61$, $p=.006$, $\eta^2=.32$). Again, and as per Prediction C, the only items to prompt significant pupil dilations were ‘homophobic’-items with disorder-inconsistent sequels ($t(20)=2.57$, $p=.018$). Also for this key group, mean maximum pupil size following offset of disorder-inconsistent items with ‘homophobic’ was significantly and marginally larger than for counterparts using ‘dislike’ ($t(20)=-2.82$, $p=.011$) or ‘prejudice’ ($t(20)=2.40$, $p=.026$), respectively.

4.3. Discussion

That disorder-inconsistent ‘homophobic’-items, and no others, prompt significant pupil dilations, and do so regardless of whether participants hold a disorder theory, strongly suggests that ‘S is homophobic’ triggers, in competent speakers quite generally, inferences to *S is mentally ill*, which are driven by a word-associated stereotype (rather than a differently anchored implicit theory). But these inferences lead to different plausibility ratings from the groups with different background beliefs: Only implicit disorder-theorists judge disorder-consistent ‘homophobic’-items more plausible than their disorder-inconsistent counterparts,

while innocent participants deem the latter more plausible. Despite our label, the 'innocents' presumably hold a view on the matter, viz., that aversion and prejudice against homosexuals labelled as 'homophobia' are *not* due to any mental disorder. Accordingly, they suppress the stereotypical conclusion at odds with their background belief. Significantly longer reaction times to 'homophobic'-items in the disorder-inconsistent than disorder-consistent condition could support this suggestion, but the lack of interaction and main effects in our reaction time data does not allow us to follow this up.

Competent speakers automatically infer that S is mentally ill, from 'S is homophobic' but not from the two other verb phrases. Even so, implicit disorder-theorists give numerically almost identical plausibility ratings for disorder-consistent items with any of the three (dislike: 3.51, homophobic: 3.49, prejudice: 3.51), as do, with duly lower values, the 'innocents' (2.65, 2.78, 2.68). This suggests these plausibility assessments are largely determined by participants' implicit theory or background beliefs. Since we would expect activation of implicit theories during comprehension to show up through pupil dilations in all three conditions, we conclude that these background beliefs are activated only during the plausibility assessment task. We lack the space to discuss the apparent tension between this conclusion and studies suggesting comprehensive bodies of general event knowledge are activated, at the earliest possible moment, during incremental language comprehension (reviews: Elman 2009, Metusalem 2012).

This study demonstrates that initial stereotypical inferences (picked up by pupillometry) need not go on to influence subsequent judgments (reflected by plausibility ratings), which may be shaped by background beliefs not associated with specific linguistic expressions. The upshot for our possible libel case is that the prosecution proceeds from a true premise but remains unsuccessful: In language comprehension, competent speakers indeed make automatic inferences from 'S is homophobic' to potentially libellous attributions of mental disorder. But whether this implication is accepted as plausible depends upon the recipient's background beliefs. Crucially, these are brought to bear with just the same result on attributions of attitudes (dislike) which typically do not qualify as defamatory, even when false (Section 2.3). At any rate in the student population sampled, the use of 'homophobic' is not inherently more libellous than talk of 'strong dislike of homosexuality' – i.e. not at all.

The more general methodological upshot is that the moment we turn from psycholinguistic questions about language comprehension processes to philosophical questions about how stereotypical inferences affect our judgments and further reasoning, we should combine online with offline measures, in a comprehensive toolkit like the one proposed.[12]

References

Alexander, J. (2012). *Experimental Philosophy*. Cambridge: Polity.

Alexander, J., Gonnerman, C., & Waterman, J. (2015). Salience, and epistemic egocentrism. In J. Beebe (ed.), *Advances in Experimental Epistemology* (pp.97-118). London:

[12] Oliver Afridijanta assisted with data collection, Mark Curtis with legal research. For helpful comments on previous drafts, we thank James Hampton, an anonymous referee, and audiences in Reading, Berlin, and Buffalo.

Bloomsbury.

Alter, A.L. & Oppenheimer, D.M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review 13*, 219-235.

Angermeyer, M.C. & Dietrich, S. (2006). Public beliefs about and attitudes towards people with mental illness: a review of population studies. *Acta Psychiatrica Scandinavica*, 113: 163–179.

Armstrong, R.A. (2014). When to use the Bonferroni correction. *Ophthalmic and Psychological Optics*, 34: 502-508.

Atlas, J. D., & Levinson, S. C. (1981). It-clefts, informativeness and logical form: Radical pragmatics (revised standard version). In P. Cole (ed.), *Radical Pragmatics* (pp. 1-62). New York: Academic Press.

Ayer, A.J. (1956). *The Problem of Knowledge*. Repr. 1990. London: Penguin.

Boyd, K. & Nagel, J. (2014). The reliability of epistemic intuitions. In E. Machery and E. O'Neill (eds.), *Current Controversies in Experimental Philosophy* (pp. 109-127). London: Routledge.

Brewer, B. (2011). *Perception and Its Objects.* Oxford: OUP.

Brown, P.M. & Dell, G.S. (1987). Adapting production to comprehension: The explicit mention of instruments. *Cognitive Psychology 19*: 441–472.

Cabin, R.J., & Mitchell, R.J. (2000). To Bonferroni or not to Bonferroni: When and how are the questions. *Bulletin of the Ecological Society of America*, 81: 246-248.

Carruthers, P. (2002). The cognitive functions of language. *Behavioral and Brain Sciences*, 25: 657–674

Carston, R. (2013). Legal texts and canons of construction: A view from current pragmatic theory. In M. Freeman & F. Smith (eds.): *Law and Language* (pp. 8-33). Oxford: OUP.

Christians, C.G., Fackler, M., Richardson, K.B., Kreshel, P.J., Woods, R.H., & Woods, R.H. (2016). *Media Ethics: Cases and Moral Reasoning*, 9th ed., London: Routledge

Crisp, A.H., Gelder, M.G., Rix, S., Meltzer, H.I., & Rowlands, O.J. (2000). Stigmatisation of people with mental illnesses. *The British Journal of Psychiatry*, 177: 4-7.

De Neys, W. (2006). Automatic-heuristic and executive-analytic processing during reasoning: chronometric and dual-task considerations. *The Quarterly Journal of Experimental Psychology*, 59: 1070-1100

Devine, P.G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology, 56:* 5-18.

Devine, P.G. (2001). Implicit prejudice and stereotyping: How automatic are they? *Journal of Personality and Social Psychology, 81:* 757-759.

Durant, A. (1996). On the interpretation of allusions and other innuendo meanings in libel actions. *International Journal of Speech Language and the Law*, 3: 195-201

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

Engelhardt, P.E., Ferreira, F., & Patsenko, E.G. (2010). Pupillometry reveals processing load

during spoken language comprehension. *Quarterly Journal of Experimental Psychology, 63*: 639-645.

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

Fein, O., Yeari, M., & Giora, R. (2015). On the priority of salience-based interpretations: the case of sarcastic irony. *Intercultural Pragmatics,* 12, 1-32.

Ferretti, T., McRae, K., & Hatherell, A. (2001). Integrating verbs, situation schemas, and thematic role concepts. *Journal of Memory and Language, 44,* 516-547.

Fischer, E. (2014). Philosophical intuitions, heuristics, and metaphors. *Synthese,* 191: 569-606

Fischer, E., & Collins, J. (2015). Rationalism and naturalism in the age of experimental philosophy. In their (eds.): *Experimental Philosophy, Rationalism and Naturalism* (pp. 3-33). London: Routledge.

Fischer, E., & Engelhardt, P.E. (2016). Intuitions' linguistic sources: Stereotypes, intuitions, and illusions. *Mind & Language,* 31, 65-101.

Fischer, E., & Engelhardt, P.E. (in press). Diagnostic Experimental Philosophy. *Teorema*

Fischer, E., Engelhardt, P.E., & Herbelot, A. (2015). Intuitions and illusions. From explanation and experiment to assessment. In: E. Fischer & J. Collins (eds.): *Experimental Philosophy, Rationalism and Naturalism.* (pp. 259-292). London: Routledge.

Friederici, A.D., & Frisch, S. (2000). Verb argument structure processing: The role of verb-specific and argument-specific information. *Journal of Memory and Language, 43:* 467-507.

Gentner, D., Imai, M. & Boroditsky, L. (2002). As time goes by: evidence for two systems in processing space-time metaphors. *Language and Cognitive Processes 17*: 537-565

Gerken, M., & Beebe, J. (2016). Knowledge in and out of contrast. *Nous,* 50: 133-164.

Gibson, E. (1998). Linguistic complexity: locality and syntactic dependencies. *Cognition, 68,* 1-76.

Gilbert, D.T (1991). How mental systems believe. *American Psychologist,* 46: 107-119

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

Giora, R., & Fein, O. (1999). On understanding familiar and less-familiar figurative language. *Journal of Pragmatics* 31, 1601-1618

Givoni, S., Giora, R., & Bergerbest, D. (2013). How speakers alert addressees to multiple meanings, *Journal of Pragmatics, 48,* 29-40

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

Hallahan, K., (2011). Political public relations and strategic framing. In J. Strömbäck & S. Kiousis (eds.): *Political public relations: principles and applications* (pp.177-213). New York: Routledge.

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

Harmon-Vukić, M., Guéraud, S., Lassonde, K.A. & O'Brien, E.J. (2009). The activation and instantiation of instrumental inferences. *Discourse Processes,46*: 467-490

Just, M.A., & Carpenter, P.A. (1977). *Cognitive processes in comprehension.* Hillsdale, NJ: Erlbaum.

Kahneman, D. (1973). *Attention and Effort.* Englewood Cliffs, N.J.: Prentice-Hall.

Kahneman, D. & Frederick, S. (2005). A model of heuristic judgment. In K.J. Holyoak & R. Morrison (eds.), *Cambridge Handbook of Thinking and Reasoning* (pp. 67-293). Cambridge: CUP

Klin, C.M., Guzman, A.E. & Levine, W.H. (1999). Prevalence and persistence of predictive inferences. *Journal of Memory and Language, 40*: 593–604.

Knobe, J. (2016). Experimental philosophy is cognitive science. In J. Sytsma & W. Buckwalter (eds.), *Blackwell Companion to Experimental Philosophy* (pp.37-52). Wiley Blackwell: Malden, MA.

Knobe, J.& Nichols, S. (2008). An experimental philosophy manifesto. In their (eds), *Experimental Philosophy* (pp. 3-14). Oxford: OUP

Koriat, A. (2012). The self-consistency model of subjective confidence. *Psychological Review, 119*: 80-113

Kutas, M. & Federmeier, K.T. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology, 62*: 621-647

Laeng, B., Sirois, S., & Gredeback, G. (2012). Pupillometry: A window to the preconscious? *Perspectives on Psychological Science, 7:* 18-27.

Levelt, W.J.M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press

Levine, T.R. (2014). Truth-default theory (TFT): A theory of human deception and deception detection. *Journal of Language and Social Psychology*, 33: 378-392

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

Loftus, E.F., Miller, D.G., & Burns, H.J. (1978). Semantic integration of verbal information into a visual memory. *Journal of Experimental Psychology: Human Learning and Memory, 4:* 19-31.

Lunney, M. & Oliphant, K. (2010). *Tort Law: Text and Materials.* Oxford: OUP.

Marmor, A. (2014). *The Language of Law*. Oxford: OUP.

McRae, K., Ferretti, T.R., & Amyote, I. (1997). Thematic roles as verb-specific concepts. *Language and Cognitive Processes, 12*: 137-176

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

Metusalem, R., Kutas, M., Urbach, T.P., Hare, M., McRae, K., Elman, J.L. (2012). Generalized event knowledge activation during online sentence comprehension. *Journal of Memory and Language*, 66, 545-567.

Morewedge, C.K. & Kahneman, D. (2010). Associative processes in intuitive judgment.

Trends in Cognitive Science, 14: 435-440

Nagel, J. (2012). Intuitions and experiments: a defence of the case method in epistemology. *Philosophy and Phenomenological Research, 85*: 495-527

Nakagawa, S. (2004). A farewell to Bonferroni: the problems of low statistical power and publication bias. *Behavioral Ecology*, 15: 1044-1045.

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education, 15:* 625-632.

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.

Osterhout, L., Bersick, M., & McLaughlin, J. (1997). Brain potentials reflect violations of gender stereotypes. *Memory & Cognition, 25:* 273-285.

Parcesepe, A. M., & Cabassa, L. J. (2013). Public stigma of mental illness in the United States: A systematic literature review. *Administration and Policy in Mental Health,* 40: 384-399

Hampton, J.A. & Passanisi, A. (2016). When intensions do not map onto extensions: Individual differences in conceptualization. *Journal of Experimental Psychology: Learning Memory and Cognition*, 42: 505-523.

Patson, N.D. & Warren, T. (2010). Eye movements to plausibility violations. *Quarterly Journal of Experimental Psychology,* 63: 1516-1532

Peel, E. & Goudcamp, J. (2014). *Winfield and Jolowicz on Tort*. London: Sweet & Maxwell.

Pickering, M.J. & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences, 36*: 329-392

Poggi, F. 2011: Law and conversational implicatures. *International Journal for the Semiotics of Law*, 24: 21–40

Pohl, R. (ed.) (2004). *Cognitive Illusions*. New York: Psychology Press

Pust, J. (2012). Intuition. In: E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, http://plato.stanford.edu/ archives/win2012/entries/intuition/

Robinson, H. (2001). *Perception*. London: Routledge

Rudman, L.A., Ashmore, R.D., & Gary, M.L. (2001). "Unlearning" automatic biases: The malleability of implicit prejudice and stereotypes. *Journal of Personality and Social Psychology, 81:* 856-868.

Rumelhart. D. E. (1978). Schemata: The building blocks of cognition. In R. Spiro, B. Bruce, & W. Brewer (eds.), *Theoretical Issues in Reading Comprehension*. Hillsdale, N.J.: Erlbaum

Sack, R.D. (2010). *Sack on Defamation: Libel, Slander, and Related Problems*, 4th ed., vol.1. New York: Practising Law Institute.

Sheehan, K.B. (2014). *Controversies in Contemporary Advertising*. Los Angeles: Sage

Shuy, R.W. (2010). *The Language of Defamation Cases*. Oxford: OUP

Simpson, G.B., & Burgess, C. (1985). Activation and selection processes in the recognition of

ambiguous words. *Journal of Experimental Psychology: Human Perception and Performance, 11*: 28-39

Smith, A.D. (2002). *The Problem of Perception*. Cambridge, Mass: Harvard UP

Soames, S. (2008). Interpreting legal texts: What is, and what is not, special about the law. In his: *Philosophical Essays, Vol. 1* (pp.403–424). Princeton: Princeton UP

Solum, L. 2013: Communicative content and legal content. *Notre Dame Law Review,* 89: 479–520

Squire, L.R. (1992). Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review, 99,* 195-231.

Stephens, G.J., Silbert, L.J., & Hasson, U. (2010). Speaker–listener neural coupling underlies successful communication. *PNAS 107*: 14425–14430

Stich, S., & Tobia, K. (2016). Experimental philosophy and the philosophical tradition. In J. Sytsma & W. Buckwalter (eds.), *Blackwell Companion to Experimental Philosophy* (pp.5-21). Wiley Blackwell: Malden.

Sturt, P. (2003). A new look at the syntax-discourse interface: The use of binding principles in sentence processing. *Journal of Psycholinguistic Research, 32,* 125-139.

Tabachnick, B.G., & Fidell, L.S. (1989). *Multivariate Statistics.* Harper and Row, New York.

Thompson, V. A., Prowse Turner, J. A. and Pennycook, G. 2011: Intuition, reason, and metacognition. *Cognitive Psychology 63*: 107-140

Till, R.E., Mross, E.F., & Kintsch, W. (1988). Time course of priming for associate and inference words in a discourse context. *Journal of Verbal Learning and Verbal Behaviour, 16,* 283-298.

Traugott, E.C. & Dasher, R.B. (2005). *Regularity in Semantic Change.* Cambridge: CUP.

Trueswell, J.C. (1996). The role of lexical frequency in syntactic ambiguity resolution. *Journal of Memory and Language*, 35: 566-585.

Weinberg, G. 1973: *Society and the healthy homosexual*. New York, NY: Doubleday.

Weinberg. J. (2015). Humans as instruments, on the inevitability of experimental philosophy. In: E. Fischer & J. Collins (eds.): *Experimental Philosophy, Rationalism, and Naturalism* (pp. 171-187). London: Routledge.

Weinberg, J. S., Nichols, S & Stich, S. (2001). Normativity and epistemic intuitions. *Philosophical Topics*, 29: 429–460.

Zwaan, R.A. (1999). Situation models: The mental leap into imagined worlds. *Current Directions in Psychological Science, 8:* 15-18.

Zwaan, R.A., & Radvansky, G.A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin, 123:* 162-185.

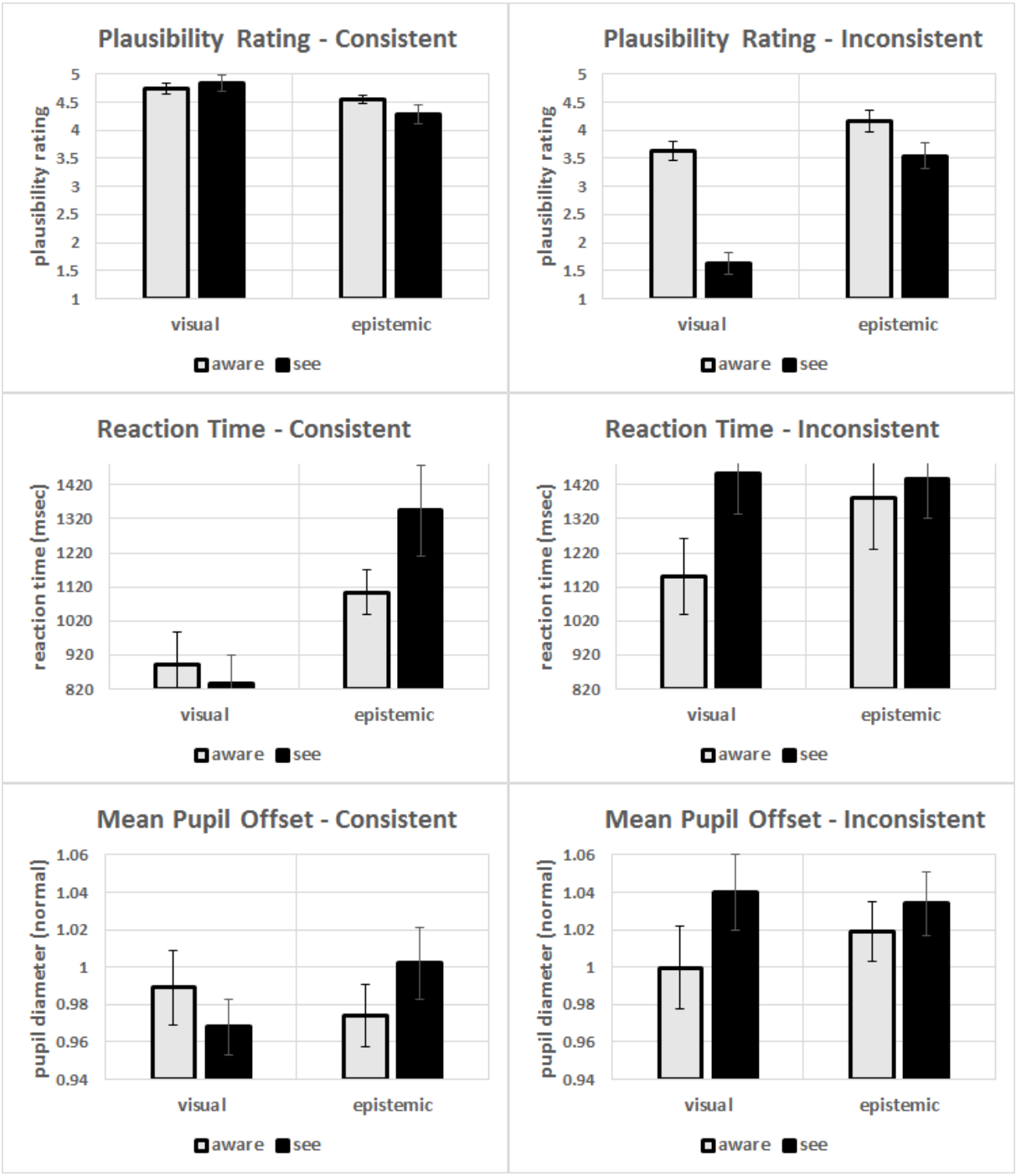

Figure 1: Top panels show mean plausibility ratings. Middle panels show reaction times. Bottom panels show mean pupil diameter during the offset time window. Left panels show the s-consistent contexts; right panels show the s-inconsistent contexts. Error bars show the standard error of the mean.

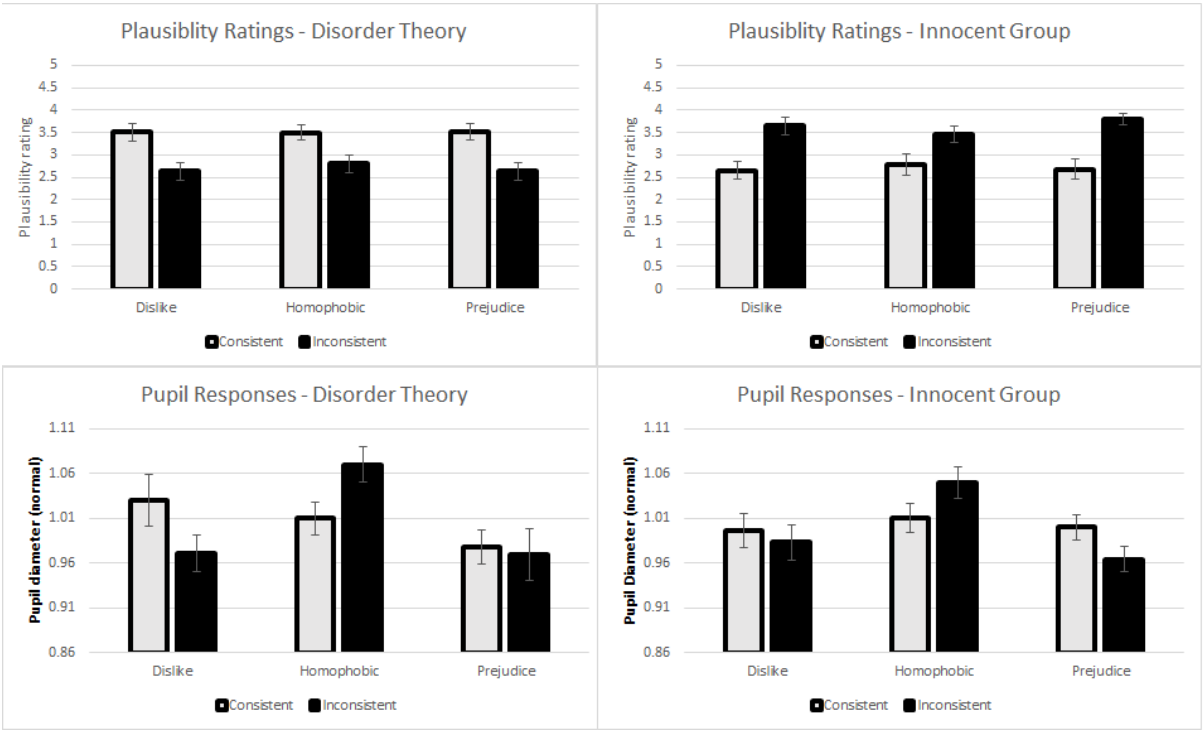

Figure 2: Top panels show mean plausibility ratings. Bottom panels show mean pupil diameter during the offset time window. Left panels show results for the disorder theory group, right panels for the innocent group. Error bars show the standard error of the mean.