Five degrees of (non)sense: Investigating the connection between bullshit receptivity and susceptibility to semantic illusions

Dario Paape (University of Potsdam)

The sentence *The invisible is beyond new timelessness* is bullshit [1]. Bullshit is characterized by *unclarifiable unclarity*: It has no clear meaning that could be explained without significant deviation from the original form of the statement [2]. Yet, bullshit statements are often judged to be true or even profound [1,3]. Individual differences in bullshit receptivity have been partly attributed to differences in acquiescence bias and/or interpretive charity [1,3]. Participants who see patterns in visual noise also tend to endorse bullshit sentences, suggesting that they tend to "creat[e] meaning where no meaning exists" [4]. Such a tendency may extend to other forms of nonsense. Semantic illusions arise in sentences such as *No head injury is too trivial to be ignored* ("depth charge" illusion) [5] or *More people have been to Russia than I have* (comparative illusion) [6], which are often perceived as sensible despite being compositionally incongruous. We have two aims: To investigate the possible correlation between individual bullshit receptivity and susceptibility to semantic illusions, and to investigate the possible shared role of interpretive charity and illusory pattern perception.

As a cover story for our experiment, we told participants that we had created an artificial intelligence (AI) system that can create natural-sounding utterances but also occasionally produces nonsense. 100 participants were asked to rate the stimulus sentences' meaningfulness and naturalness on a 7-point Likert scale. In addition to bullshit statements and semantic illusion sentences, we included sensible sentences and transparently nonsensical sentences as controls, as shown in **Table 1**. As a measure of illusory pattern perception, participants were shown randomly generated two-dimensional dot patterns (see **Figure 1**) and told that these represented the AI's neuronal activations. They were told to indicate for each pattern whether they saw any meaningful structure in the activations or not. Reaction times were collected for both tasks.

Ratings were analyzed with a hierarchical cumulative logit model. To control for differences in the use of the Likert scale, the model contained subject-specific adjustments (random effects) to the sizes of the rating "bins". Crucially, we estimated the correlations between subject-specific adjustments to the mean ratings — relative to the average — across sentence types, as well as between ratings and the pattern perception measure. For example, if perceived meaningfulness is due to interpretive charity, the subject-wise adjustments across sentence types should be positively correlated, as charitable subjects should give higher ratings across the board.

Our results indicate that interpretive charity plays a role in bullshit receptivity: Participants who gave higher ratings to nonsense and sensible sentences also gave higher ratings to bullshit sentences (**Figure 2**). Furthermore, there is a negative correlation between ratings for nonsense and sensible sentences, due to subjects being more or less extreme in their negative perception of nonsense and their positive perception of sensible sentences. Participants who made stronger distinctions showed a stronger effect of sentence length on reading time, suggesting that they read more attentively [7]. There is no indication of a correlation between bullshit receptivity and susceptibility to semantic illusions, nor of a general correlation with illusory pattern perception. Only the comparative illusion shows some indication of a positive correlation with pattern perception, as well as of a negative correlation with attentive reading. By contrast, the depth charge illusion shows some indication of a positive correlation with attentive reading, and of a negative correlation with nonsense acceptability. Overall, our results suggest that there may be no general individual trait that explains bullshit receptivity and susceptibility to semantic illusions. However, the results raise interesting possibilities for future research: The depth charge and comparative illusions may involve different cognitive mechanisms, and may be differentially related to attention and depth of processing.

(1) (a) **Sensible**
Your teacher can open the door, but you must enter by yourself.

(b) **Nonsense**
One can say that flowers with a lot of old nettles do not limp without great experience value.

(c) **Bullshit**
The invisible is beyond new timelessness.

(d) **Dept charge illusion**
No head injury is too trivial to be ignored. (correct: . . . to be treated)

(e) **Comparative illusion**
More people have been to Russia than I have.

Table 1. Example sentences used in the experiment. Half of the sensible sentences were "profound" like the example, while the other half were more mundane (*Newborns need constant attention*). The bullshit condition contained an equal number of "pseudo-profound" bullshit [1], scientific bullshit [3], and "International Art English" [8] sentences.

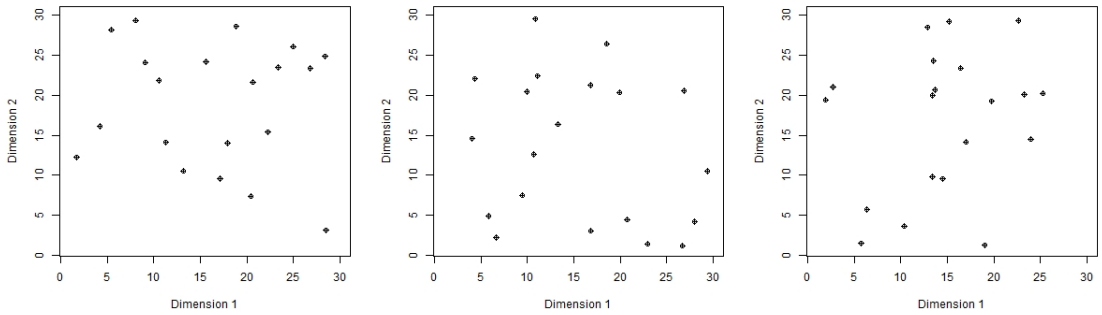

Figure 1. Example dot patterns ("neuronal activations") used in the pattern recognition task.

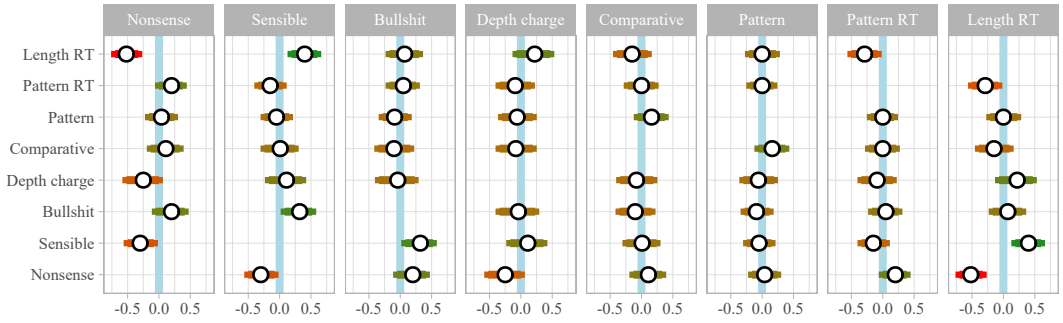

Figure 2. Estimates and 95% credible intervals of subject-level random-effects correlations.

References. [1] Pennycook et al. (2015, Judg Dec Mak). [2] Cohen (2002, *Deeper into bullshit*). [3] Evans et al. (2020, Judg Dec Mak). [4] Walker et al. (2019, Judg Dec Mak). [5] Wason & Reich (1979, Q J Exp Psychol). [6] Wellwood et al. (2018, J Semant). [7] Schad et al. (2012, Cognition). [8] Turpin et al. (2019, Judg Dec Mak).