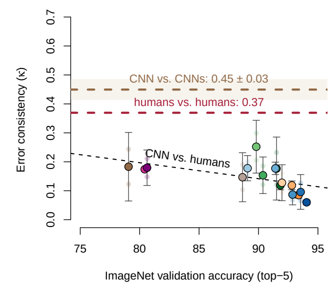
We would like to thank all reviewers for their valuable feedback and we very much appreciate their assessment of our work as *extremely relevant to the NeurIPS community* and *extremely well written* (**R1**), a *principled evaluation* and potentially a *highly impactful paper* (**R2**) with *novel* and *very intriguing findings* (**R3**). Furthermore, all reviewers were confident that our work can be reproduced, and pointed out how it could *encourage the field to explore a wider space of architectures and training schemes* (**R2**) and *attend the consistency of behaviour rather than aggregate measures* (**R4**).



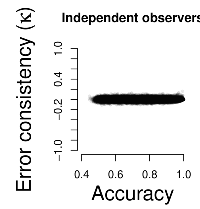
R1, **R2**, **R3**, **R4**: *Discussion of Brain-Score / CORnet is overly critical. Find way of unifying benchmarks.* We apologise for our overly critical presentation of Brain-Score and CORnet. Together with CORnet/Brain-Score authors Kubilius and Schrimpf we re-phrased numerous unfair or misleading statements and now have a balanced manuscript we and K&S all agree upon. K&S believe error consistency to be an important behavioural metric and want to include it on Brain-Score.

R1: *CNNs are not trained on stimuli* / **R2**: *Repeat experiment on dataset where CNNs and humans have similar performance* / **R3**, **R4**: *Repeat experiment with natural ImageNet images as baseline.* We now include standard ImageNet images where human and pre-trained CNN accuracies are both very high and similar ($.960 \pm .036\%$). New results included in the paper (shown on the right) complement previous findings. Thanks for this excellent suggestion which makes the paper stronger!

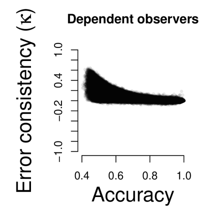
R1: *Clarify term "strategy".* We now clarify the difference between "high-level strategy" and "decision rule" (for decision rule: following the terminology from "Shortcut learning in deep neural networks", GEIRHOS ET AL, 2020).

R1: *Only CNNs trained on ImageNet were used* / **R2**: *Include models trained on Stylized-ImageNet.* Again an excellent suggestion which we have incorporated into our final manuscript. We analyzed three CNNs with different degrees of stylized training data. Model shape bias predicts human-CNN error consistency for cue conflict stimuli, indicating that networks basing their decisions on object shape (rather than texture) make more human-like errors:

model shape bias (%)	20.5	21.4	34.7	81.4
human-CNN consistency (κ)	.066	.068	.098	.195



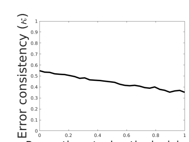
R2: *Does kappa really disentangle error consistency from accuracy?* In Figure 2b, κ and c_{exp} are not correlated (r=-0.00015, $p > 0.05$); a simulation (see right plot) confirms: κ and accuracy are not correlated for independent decision makers (r=-0.004, $p > 0.05$). For *dependent* observers, any pattern is possible: zero correlation (Figures 3a, 3b), positive correlation (Figure 3c) and even negative correlation (simulated toy experiment, bottom figure on the right). Thus crucially, there is *no* correlation between consistency (κ) and accuracy for independent observers whilst for dependent (consistent) observers correlations are possible but they are a property of the decision makers, not the analysis. We now discuss this point in the main paper and show the simulations in the appendix.



R3: *A closer analysis of error differences would be helpful. / Detailed comparison to Brain-Score.* Another nice suggestion! We now visualize striking error differences between CNNs and humans for all experiments (original ImageNet images, cue conflict, silhouette, edges) and discuss potential underlying causes. Example visualized below. Top row: "Hard" images for CNNs (correctly classified by all humans but not by *any* CNN). Bottom row: "Hard" images for humans (`bear`, `bear`, `bird`, `oven`). Additionally, we now plot confusion matrices to analyse category-level errors. Concerning comparison to Brain-Score conditions V1, V2, V4, IT, behaviour: This is already done in the appendix (SF.7, SF.8, SF.9); now linked & discussed more prominently.

R4: *The main problem of the paper is the use of unnatural stimuli for testing.* First we now include "natural" ImageNet images leading to the very same conclusions as with "unnatural" stimuli (see above). Second, we strongly believe in the value of investigating model behaviour with controlled "unnatural" stimuli: Significant progress in neuroscience—e.g. discovering receptive fields of simple and complex cells—was made using "unnatural" bar-like stimuli. In deep learning adversarial examples and texture bias were discovered by testing models on (unnatural) images different than the training data. Clearly we can learn a lot about the inner workings of a system by probing it with *appropriate* artificial stimuli ("In praise of artifice", RUST & MOVSHON, 2005; "In praise of artifice reloaded", MARTINEZ-GARCIA ET AL, 2019); we now state this motivation more explicitly.

R4: *Aggregating the classification probability by arithmetic mean may not be optimal.* We have now included a principled derivation showing that the arithmetic mean is, perhaps counterintuitively, optimal. Essentially, the aggregation can be derived by calculating the posterior distribution $p(c|x)$ of a discriminatively trained CNN under a new prior chosen at test time (here: $\frac{1}{16}$ over coarse classes); resulting in decision $C|x = \operatorname{argmax}_C \sum_{c \in C} \frac{1}{|C|} p(c|x)$.



R4: *Other suggestions.* (1) We now state why we did not include a comparison of κ to the (Pearson) correlation coefficient since the literature rejects correlation as a measure of agreement (HUNT 1986, WATSON & PETRIE, 2010). (2) We used deterministic decisions throughout the paper. When a proportion of decisions are stochastically sampled from the softmax instead, consistency between two CNNs decreases slightly (see plot on the right).