Title: Stimulus models test hypotheses in brains and DNNs

Philippe G. Schyns(*), Lukas Snoek & Christoph Daube, School of Psychology and Neuroscience, University of Glasgow

Correspondence: philippe.schyns@glasgow.ac.uk

Keywords: Generative models, perception, experimental design,

Pitkow [1] raises several points about our "Degrees of equivalence between the brain and its DNN models" *Opinion Article* [2]. Here, we clarify our position before delving into the elusive argument of the "suitably rich" stimulus ensemble that Pitkow suggests as an alternative to our 2nd degree of algorithmic equivalence (see also section *Is the 2nd degree really necessary?* in our *Opinion Article* [2]).

Remember that we seek to explain the inaccessible algorithms of the brain using the more accessible DNN models. At their 1st degree of equivalence, we follow current practice, ask whether DNN models can predict responses, and compare Stimulus-Response relationships $\langle S, R \rangle$ across systems, necessarily using a *finite* sample of stimuli *S*–effectively only very few when imaging the brain.

Pitkow's argument refers to the practical problem of selecting this finite *S*. To do so, he wisely recommends that we should not "[lose] sight of fundamental challenge for intelligence: generalization to new things." And, most importantly, that the *utility* of *S* is defined by its *diversity* and *relevance to the tuning of the system*. For example, a set of 2D images of trees would likely have low utility to test a system tuned to categorize the 3D shapes of noses in faces.

Though we can only agree with such general recommendations (see **Box 1**), how will we resolve the practical problem? That is, how can we select *S* with the desired utility? Our solution samples a generative model of the stimulus, whose features (*F*) are formal hypotheses about the tuning of the brain. At the 2nd degree of algorithmic equivalence, we can then effectively test whether the brain and DNNs produce the same *R* to these same generated *S*, *because*, within the limits of what the generative model can decorrelate, both process the same categorization features F–in the example, the F that generate differently shaped 3D noses in faces. That is, we test $\langle S, R \rangle_F$ equivalence.

Importantly, the 2nd degree $\langle S, R \rangle_F$ is relative to the hypothesized F that the stimulus model can control and decorrelate. Thus, we should test the 2nd degree using different, explicit stimulus models (i.e. different hypotheses on *F*), just as we test the 1st degree using different DNN models of the $\langle S, R \rangle$ relationship. The key advantage of sampling from stimulus models over sampling from 2D images of databases is that the utility of *S* (cf. its diversity and relevance to the tuning of the system) becomes an optimizable cost function that depends on the hypothesized F, rather than solely on the amount of naturalistic images, as Pitkow suggests.

To illustrate the pivotal role of the 2nd degree, we used the counterexample of an impoverished stimulus sample *S*, where the eyes and mouths of faces are correlated, leading to identical "face" responses across systems that process either the eyes or the mouths. Pitkow suggests increasing the utility of *S* with stimuli that decorrelate eyes and mouths. This suggestion is indeed at the centre of the hypothesis-testing epistemology of our pivotal 2nd degree, where F-hypotheses (here, the brain is tuned to eyes and mouths) are required to quantify the "suitable richness" of *S*, and are, in fact, the first step to formalize

generative models of the stimulus. This formalization then enables systematic testing of the psychophysical generalization gradients of systems along the generative F–e.g., controlled changes of only 3D generative F of shape, illumination, orientation, shapes of eyes, noses and mouths, etc., and their interactions (see section *Generalization Gradients* of our *Opinion Article [2]*). In contrast, blind sampling of uncharacterized naturalistic images might never deliver a "suitably rich" *S* [3–5].

In sum, Pitkow's "suitably rich" *S* remains an elusive (and experimentally impractical in neuroimaging) black-box compared to the *S* gathered from sampling formal generative models of 3D stimuli–where different models with different generative *F* can also compete to account for the 2nd degree $\langle S, R \rangle_F$. Psychophysics modelled the generative process of simple luminance contrasts with sinewave gratings a long time ago. Neuroimaging of visual cognition and its DNN modelling should embrace this approach with timely generative models of complex 3D faces, bodies, objects and scenes (see [6] for the same argument coming from DNN modelling). These are now within grasp [7,8].

To complete the argument, when the brain and DNN models are equated at the 2nd degree of equivalent categorization features (i.e. satisfy the $\langle S, R \rangle_F$ relationship), it becomes meaningful to compare across systems the algorithmic computations that process these same features at the 3rd degree of equivalence.

Box 1

Are naturalistic stimuli more important than synthetic stimuli?

We agree with Pitkow that the "tuning relevance" and "richness" of *S* affect the conclusions that can be drawn from an experiment using *S*. We also agree that covering a "wide range of naturalistic stimuli" is an interesting starting point to explore the $\langle S, R \rangle$ relationship, to allow for a distribution of responses related to the distribution stimuli [9]. However, Pitkow's conclusion, that sampling naturalistic stimuli is "more important," reflects the fallacy at the core of adversarial examples [10], which occur in locations in stimulus space that were not part of the naturalistic training set. Counterintuitive DNN predictions to adversarial examples reveal counterintuitive algorithmic differences between humans and DNN models, that remain hidden when we only test with naturalistic stimuli. Underappreciating such non-naturalistic locations in stimulus space can therefore overestimate the generalization capabilities of a DNN model (cf. the "complete functional description" [11]). Taken together, exploratory approaches relying on naturalistic stimuli and testing the specific hypotheses of generative stimulus features are both useful for their respective purposes [12]. Thus, we disagree that one is in principle "more important" than the other. We however also believe that vision sciences have matured enough for the testing of specific hypotheses.

1. Pitkow, Xaq (2022) Algorithmic similarity depends continuously on the input distribution, not categorically on how inputs are generated. *Trends Cogn. Sci.*
2. Schyns, P.G. *et al.* (2022) Degrees of algorithmic equivalence between the brain and its DNN models. *Trends Cogn. Sci.* 26, 1090–1102
3. Golan, T. *et al.* (2020) Controversial stimuli: Pitting neural networks against each other as models of human cognition. *Proc. Natl. Acad. Sci.* 117, 29330–29337
4. Golan, T. *et al.* (2022) Distinguishing representational geometries with controversial stimuli: Bayesian experimental design and its application to face dissimilarity judgments. presented at the SVRHM 2022 Workshop @ NeurIPS
5. Bashivan, P. *et al.* (2019) Neural population control via deep image synthesis. *Science* 364, eaav9436

6. de Melo, C.M. *et al.* (2022) Next-generation deep learning based on simulators and synthetic data. *Trends Cogn. Sci.* 26, 174–187
7. Gan, C. *et al.* (2021) ThreeDWorld: A Platform for Interactive Multi-Modal Physical Simulation. *ArXiv200704954 Cs* at <http://arxiv.org/abs/2007.04954>
8. Henderson, P. *et al.* (2020) Leveraging 2d data to learn textured 3d mesh generation. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7498–7507
9. Rieke, F. *et al.* (1995) Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents. *Proc. R. Soc. Lond. B Biol. Sci.* 262, 259–265
10. Szegedy, C. *et al.* (2014) Intriguing properties of neural networks. *ArXiv13126199 Cs* at <http://arxiv.org/abs/1312.6199>
11. Naselaris, T. *et al.* (2011) Encoding and decoding in fMRI. *NeuroImage* 56, 400–410
12. Tukey, J.W. (1980) We Need Both Exploratory and Confirmatory. *Am. Stat.* 34, 23–25