Deep neural networks (DNNs) have become powerful and increasingly ubiquitous tools to model human cognition, and often produce similar behaviors. For example, with their hierarchical, brain-inspired organization of computations, DNNs apparently categorize real-world images in the same way as humans do. Does this imply that their categorization algorithms are also similar? We have framed the question with three embedded degrees that progressively constrain algorithmic similarity evaluations: equivalence of (i) behavioral/brain responses, which is current practice, (ii) the stimulus features that are processed to produce these outcomes, which is more constraining, and (iii) the algorithms that process these shared features, the ultimate goal. To improve DNNs as models of cognition, we develop for each degree an increasingly constrained benchmark that specifies the epistemological conditions for the considered equivalence.

**DNN models of human cognition**

As is often said, models in science should be used as simple, but not simpler, mechanisms to formulate less-accessible systems – by being glass boxes that model black boxes. In cognitive science, models offer explicit hypotheses for the computations (see Glossary) that realize a variety of cognitive abilities [1–4]. Developing from connectionism, deep neural networks (DNNs) have become an increasingly popular class of such algorithmic models [5,6] – either on their own or as front-ends to hybrid models. When using DNNs as models of human visual categorization [7–9], their brain-like layered architectures (feedforward or recurrent) perform a series of nonlinear computations that take in 2D images as inputs (e.g., of real-world faces, objects, and scenes) and deliver outputs such as their category memberships (e.g., ‘Mary’, ‘cars and buildings’, or ‘New York City’) or their mappings into low-dimensional semantics [10]. Following training, the output of a DNN can therefore be directly compared to human behavioral responses. This generally frames DNN categorizations as an end-to-end algorithmic function \( A \) that returns responses \( R \) to stimulus \( S \) (Equation 1).

\[
R = A(S)
\]  

Such end-to-end, 2D image-to-category labeling is a breakthrough in modeling history that naturally invites addressing again the enduring question [3,6,11–13]: is the algorithmic function of categorization \( A(S) \) ‘the same’ in the brain and its DNN models?

The question of ‘the same’ algorithm is notoriously difficult to frame and evaluate for the brain because we cannot directly access its unknown categorization algorithm \( A_{\text{Brain}} \), as we can in principle access that of the DNN, \( A_{\text{DNN}} \). To clarify algorithmic similarity, consider that \( A_{\text{DNN}}(S) \) is the classic feedforward composition of \( L \) embedded stages of consecutive computations [14], where each stage corresponds to one of \( L \) network layers, resulting in \( L \) stages of embedded nonlinear functions \( a \), i.e. \( a^L(…(a^2(a^1(S)))) \). The question of algorithmic similarity then becomes...
Whether, at some level of abstraction useful for a cognitive theory [3,6,12,15], the $L$ embedded computations of $A_{DNN}$ approximate those of the brain algorithm $A_{Brain}$.

Having simply restated the enduring question (leaving aside many caveats to which we return later), research practice addresses algorithmic similarity by comparing how $A_{Brain}$ and $A_{DNN}$ perform the same cognitive task. We develop here a new epistemology to develop this comparison, that models the stimuli as much as it models the responses to the stimuli, to understand how $A_{Brain}$ and $A_{DNN}$ process features of the same stimulus space, thereby introducing three embedded degrees of increasingly constrained algorithmic similarity (Figure 1, Key figure).

At this juncture it is important to clarify the scope of our contribution. Models in science are generally used in three complementary ways: to predict new data, to explain their causes, and, from this better understanding, to explore new models [16]. DNN models of cognition can predict human behavior with high accuracy, although we still need to explain how $A_{DNN}$ Causes behavior, possibly with the same computations as $A_{Brain}$. We lay out the knowledge and the conditions (i.e., the epistemology) that are pivotal to developing new DNN models of cognition, from predictions of behavior to algorithmic explanations of their causes. The exposition is therefore conceptual rather than technical, although we illustrate key points with concrete examples.

**Three increasingly constrained degrees of algorithmic equivalence**

Current practices primarily target a first degree of equivalent behavioral (or layer) outcomes between brain $A_{Brain}$ and $A_{DNN}$ [17] – in other words, the extent to which their responses ($R$) to the same stimuli are similar ($R_{DNN} \equiv R_{Human}$). A more constrained second degree requires that $A_{Brain}$ and $A_{DNN}$ process the same stimulus features ($F$) to produce these same responses [18]. Feature equivalence is pivotal to our epistemology, and allows meaningful comparisons at the third degree of the ‘how’ question, which requires that $A_{Brain}$ and $A_{DNN}$ compute this same $F$ in the same way, with the same algorithm – in other words, with the same sequence of L layered computations (cf Figure 1, where $\text{DNN feature } F_{DNN}^{1} \equiv \text{brain feature } F_{Brain}^{1}$, $F_{DNN}^{2} \equiv F_{Brain}^{2}$, ..., and $F_{DNN}^{L} \equiv F_{Brain}^{L}$). Comparisons of several DNNs at these three degrees could reveal that, of those tested, $A_{DNN}$ is most algorithmically similar to $A_{Brain}$ for the cognitive task considered, and $A_{DNN} \approx A_{Brain}$. Of course, third-degree algorithmic equivalence applies to any DNN architecture that models human cognition (e.g., feedforward and recurrent). For simplicity of exposition, we illustrate our points with the simple feedforward architecture.

Before developing the three degrees, it is useful to define an algorithm as a method formulated by a sequence of operations (basic and control structures) to resolve a task (e.g., labeling an input image) in finite time. Equivalence of algorithms between humans and their models has been the enduring goal of cognitive science [3,12], as Newell’s famous ‘20 questions with nature’ [19] highlighted half a century ago. In context, the advent of ‘end-to-end’ models recently enabled comparisons of humans and DNN algorithms that perform the same real-world categorizations. However, Newell’s injunctions for the study of human algorithms [19] – listed below, with our highlights – remain very much timely for psychology, neuroimaging, and DNN sciences alike. We will refer to these injunctions as we develop the three degrees.

‘Fundamental fact. [Human] behavior is programmable [via an information-processing algorithm].’

Injunction 1. Know the method [information-processing algorithm] that your subject is using to perform the experimental task [categorize the visual input].

Injunction 2. Never average over methods [information-processing algorithms].’

---

**Glossary**

- **Algorithm**: a guided sequence of computations (i.e., a method) that resolves a task (here, cognitive) from its inputs in finite time by producing an output (e.g., producing the category label of an input image in n steps of computations).
- **Brain feature**: a measurable property of brain responses to stimuli.
- **Computation**: the transformation of a representation.
- **Deep neural network (DNN)**: a hierarchically organized connectionist model loosely inspired by neurobiology, and whose nonlinear layers (usually many) comprise parameters (weights) that can be adjusted (‘trained’) to minimize a cost-function.
- **DNN feature**: a measurable property of DNN responses to stimuli.
- **End-to-end**: a model that learns to map raw inputs (e.g., pixel images) to a target format (e.g., category labels).
- **Generalization gradient**: the relationship between different values of the generative features of stimuli and the responses (behavioral, brain, DNN layer) that these values elicit.
- **Generative adversarial network (GAN)**: a type of generative DNN that synthesizes artificial (visual) stimuli with complex statistical structures.
- **Generative model**: a model that defines the causal components and computations to produce a defined class of outputs (e.g., images for the output class of visual stimuli).
- **Hybrid model**: a model that combines different artificial intelligence (AI) techniques, such as a DNN front-end, to label an image used for reasoning within a system of symbolic knowledge structures.
- **ImageNet**: an influential image dataset with >14 million images from the 20 000 categories that were famously used to train end-to-end categorization DNNs.
- **Layer**: a collection of artificial neurons that share the same hierarchical level within a DNN. These neurons perform computations on their inputs as specified by their weights and nonlinearities (e.g., convolution, non-linear scaling).
- **Multiple realizability**: the notion that ‘something’ (e.g., flying, adding) can be realized in multiple ways to achieve the same outcomes (e.g., wings and feathers vs. man-made articulated wings; an abacus vs. a digital computer).
First-degree equivalence of categorization behaviors
Suppose that a DNN model trained on a large image dataset (e.g., ImageNet [20]) matches trial-by-trial human categorization responses ($R_{\text{DNN}} \equiv R_{\text{Human}}$) using the $(S, R)$ stimulus–response benchmark. Does this qualify as algorithmic equivalence $A_{\text{DNN}} \equiv A_{\text{Brain}}$? At the level of behavioral outputs, yes, but only in this task [21,22] and with the stimulus distribution implicitly captured in the benchmark. Box 1 discusses how the contents of $(S, R)$ should capture the specific competence of the human categorizer in their niche, and the ensuing risks for modeling categorization when $(S, R)$ contents assume universality (as is the norm) and do not capture the niche.

Second-degree equivalence of the features that predict behavior
Suppose that an improved $(S, R)$ benchmark addresses the risks discussed in Box 1 and delivers a first-degree equivalence of categorization behaviors between $A_{\text{Brain}}$ and $A_{\text{DNN}}$. Have we now established $A_{\text{Brain}} \equiv A_{\text{DNN}}$ algorithmic equivalence? No, because we still face the classic argument of the multiple realizability [38] of responses. To illustrate with a behavioral example, Mary can categorize a face image as ‘happy’ using its mouth, whereas Peter categorizes the same face as ‘happy’ using the wrinkles around the eyes [30,31,33]. Mary and Peter similarly respond ‘happy’, although their individual algorithms are de facto different because each primarily uses a different feature from the same stimulus (i.e., $F_{\text{mouth}}$ vs. $F_{\text{eyes}}$). An identical argument applies to brain activations (e.g., measured in the right fusiform gyrus). Suppose that $A_{\text{Brain}}$ responds to images of houses, cars, and faces by processing their chimney, windshield, and eye features, whereas $A_{\text{DNN}}$ predicts features from the same stimulus.

**Key figure**

Three degrees of increasingly constrained algorithmic equivalence

---

**Figure 1.** First degree: typical equivalence (white intersection) tested between human responses (green) and 2D images of real-world 3D faces, bodies, objects, and scenes, and deep neural network (DNN) predictions (red). Such images typically form an experimental ‘black box’ because we cannot control and therefore explicitly test how real-world 3D features (projected on these 2D images) cause similar responses between humans and DNN models (white intersection). Second degree: a stimulus model generates 2D images from variations of its generative features $F$ (blue), and provides a coordinate system to chart and systematically explore the match between human and models. Such generative psychophysics controls the images, thereby enabling decomposition of human and DNN responses into: (A) the white triple intersection of similar human and DNN responses to the same controlled 3D $F$, (B) the cyan remainder of human responses to stimulus features $F$ that the DNN model cannot predict, (C) the magenta remainder of DNN predictions from $F$ that are dissimilar to human responses, and (D) the crucial yellow remainder flagging that the $F$ of this particular stimulus model fails to account for the total similarity of human and DNN behavior (suggesting testing of other $F$ from other generative stimulus models). Third degree: equivalence of the brain and DNN algorithms (across schematized brain and DNN layers) whose similar computations process the same stimulus features $F$ to produce the same responses.
At the second degree, to address the argument of multiple realizability that affects the models. Stimulus features (we add the constraint that the same responses must arise from processing the same explicit the same stimulus features. Thus, the more constrained second-degree benchmark both process the chimney to ensure that their similar activations refer to algorithms that process

Importantly, note that the second degree still does not commit to how these causal features are processed (i.e., with which algorithmic method), although of course constraining any algorithm to process specific causal features is a strong constraint on what this algorithm must do [43]. Note also that F can be idiosyncratic to the individual (Box 1), and we should not average across human...
algorithms, as is often the norm in neuroimaging studies (cf Newell’s third injunction [19,44]). Finally, it is important to emphasize that the second degree is an overarching systems constraint that is pivotal to our epistemology. When the second degree is established, we know that the brain and its DNN models use the same $F$ for categorization, and we can meaningfully open the black boxes to compare how they do so. The $F$ then become the needles whose algorithmic processing we seek to equate across the haystacks of neural and DNN model activity. How can we characterize $F$?

**Characterizing categorization features in humans and their DNN models**

The second-degree equivalence requires that we characterize what stimulus features cause responses. This is rarely achieved because the stimulus itself is not modeled. When we use 2D image databases, we stimulate DNN algorithms with pixels without controlling the process that generates these pixelated images. Naturalistic images are observational datasets without a known causal structure [45]. To understand the causal stimulus features ($F_S$) that influence categorization responses (i.e., $F_S = F_{\text{Human}}$), we need to break through this ‘wall of pixels’ and generate (i.e., control and vary) the 2D images with actual 3D generative models of face, body, object, and scene categories ([33]; also [46–49]) whose 2D projections produce the pixelated images of the database (Box 2).

Such generative modeling is the hallmark of psychophysics, where for example classic sinewave gratings are generated with four dimensions of spatial frequency, amplitude, phase, and orientation, to ascertain how each causally contributes to responses. It is time to develop an analogous generative psychophysics for higher-level vision by similarly controlling and varying the generative dimensions of dynamic 3D models of conspecifics and their surroundings [48,50]. For example, a recent study [18] used a generative model that synthesized, varied, and controlled the 3D features of faces and projected them as 2D stimuli (Box 2), which each participant rated according to their resemblance to the faces of familiar work colleagues [33]. With the same dataset, the authors independently modeled the responses of each individual participant with competing DNNs. Applying the more stringent second-order benchmark $(S, R)_{\text{F}}$ for model selection, they selected the DNN that similarly identified familiar faces because its algorithm processed similar 3D face-shape features as those provided by the generative stimulus model (i.e., $F_S = F_{\text{Human}} = F_{\text{DNN}}$).

A second-order equivalence of the modeled stimulus features addresses the multiple equivalence argument raised at the first degree – namely that multiple different features can realize the same behaviors in the human algorithm and DNN models. The study [18] could therefore conclude that $A_{\text{Brain}}$ and $A_{\text{DNN}}$ were most algorithmically similar according to the second-degree $(S, R)_{\text{F}}$-benchmark.

**Generalization gradients**

To identify the categorization features at the second degree (i.e., $F$), we must circumscribe the generative feature subset that causes similar responses. Using this subset, we can then rigorously test and compare the generalization gradients of human and DNN responses. Such generalization testing can address the notorious adversarial examples of DNNs by not only revealing catastrophic failures [63,64] but also the specific parametric range under which they occur, which is a prerequisite for developing targeted remediations. Establishing equivalent $F$ and generalization gradients at the second degree, therefore, develops more comprehensive causal explanations of human and DNN behaviors [18] with matched boundary conditions. We believe that generative psychophysics would benefit visual cognition, neuroscience, and their DNN modeling alike [48].

However, widespread adoption of generative psychophysics requires that we resolve a thorny methodological challenge. How can we parameterize the high-dimensional features of real-
Box 2. Assessing the second degree of feature equivalence \( (S, R)^F \)

In a reverse correlation experiment [30,33], a human participant sees 2D images of 3D faces and categorizes them (e.g., according to their similarity to a familiar face; black arrow in Figure 1). Different DNNs predict these human categorizations from the same 2D images (colored arrows). To establish the second degree of feature equivalence between humans and their DNN models (blue arrows), we can use classic techniques that quantify the similarities of image pixel features (i.e., heat-mapping techniques [55,56]). For example, (2) in Figure 1 illustrates locally linear receptive fields applied at the pixel level [53,54]. This technique rewrites the computations of a full non-linear DNN as a single linear weight-map per image, such that multiplying this linear weight-map with the image produces the same response as the full non-linear DNN. This is analogous to psychophysical reverse-correlation methods that quantify task-relevant pixels/decision templates in individual human participants [51,57–59]. A high correlation of such pixel maps between humans and their DNN models establishes a second degree of feature equivalence at the level of pixels [60].

However, real-world 3D faces (and bodies, objects, and scenes) hide behind the wall of their 2D pixel projections (as illustrated in Figure 1). Our cognitive explanations should break through this wall of pixels to explain human and DNN categorizations in terms of the causes of the pixels. For example, we can compute 3D shape receptive fields to represent the 3D \( F \) that explain the common causes of human (Figure I-3) and DNN responses (Figure I-4). With extensive psychophysical testing of generalization of responses (e.g., to changes of 3D shape, and also to \( F \) of illumination, scale, rotation, gender, identity, expression, and so forth), we can better test the gradient of generalization to establish the equivalence of the causal 3D \( F \).

Then, to evaluate DNN modeling results, we consider the first degree of equivalence of human and DNN responses (Figure I-5, y axis). We expand to the second degree by quantifying the similarity of the 3D \( F \) underlying human and DNN responses (e.g., by correlating the 3D receptive fields [18] or using information-theoretic redundancy [61,62]). The relationship between the two axes (Figure I-5) shows that the tested DNNs improve their first degree of behavioral predictions (y axis) when they infer the 3D \( F \) of the generative model from the 2D stimuli (x axis). Importantly, these results can then be subjected to extensive generalization testing [18].

Figure 1. Second equivalence \( (S, R)^F \). (1) Reverse correlation techniques such as Bubbles [48,51] can quantify the 2D pixel features \( F \) used by each human categorizer [52]. (2) Relatedly, visualization techniques [53,54] can quantify the 2D pixel features that cause DNN model responses. (3,4) A 3D face model [50] generates variations in the causal 3D \( F \), which project onto the wall of pixels as 2D images. (5) At the first degree, we compare how competing DNN models predict human behavior (y axis). At the second degree, we quantify the overlap between predictions from the DNNs and categorizations from the 3D \( F \) (with information-theoretic redundancy, x axis). The scatter shows that the first-degree similarity of human and DNN responses can be explained with the same causal 3D \( F \) of the generative model (second degree). Abbreviations: DNN, deep neural network; MI, mutual information; PCA, principal components analysis; RGB, red green blue.

Trends in Cognitive Sciences, December 2022, Vol. 26, No. 121095
world 3D face, object, and scene categories for realistic stimulus synthesis? From the outset we should be aware that no currently existing model might correspond to the ground truth features processed in the brain and DNN models, perhaps because the mathematics and/or visualizations necessary to explore these high-dimensional spaces are not yet sufficiently developed. To explain the stimulus causes of responses, all models might be wrong, but some could be useful. At the second degree, we should therefore compare different generative models of the stimulus in the same way as we compare different DNN models of human responses at the first degree. Two generative approaches are currently promising: direct engineering of computer graphics and indirect engineering of generative DNNs. Computer graphics can render increasingly realistic images [46,50,65]. Importantly, the parameters of 3D modeling and rendering can serve as models of the causal structure (i.e., the 3D F) of the generated 2D image categories that serve as stimuli [18,66,67]. However, such direct engineering of the generative features restricts the models of each stimulus category to the imagination of human engineers. A promising alternative uses generative adversarial networks (GANs) [68] which can produce arguably more realistic images. However, although such indirect engineering could expand generative parameters beyond human imagination, GANs are more difficult to control and therefore so far cannot deliver controlled 3D models of stimulus categories as computer graphics does (but note progress on structuring latent spaces [69–71]). Although still very much in its infancy, generative modeling of the stimuli should become as important as modeling the behavioral and brain responses to the stimuli because we cannot analyze what we cannot synthesize.

Transition: behaviorist (first degree) to pivotal cognitivist explanation (second and third degrees)

At this juncture, cognitivists would remark that the first-degree, \( \langle S, R \rangle \) equivalence is akin to stimulus–response behaviorism. The cognitive revolution expanded this restriction to the intervening internal states (here, of processing \( F \)) that enable causal explanations via successive stages of computations that relate stimulus to behavior (i.e., \( A_{\text{Brain}} \) and \( A_{\text{DNN}} \) similarly process \( F \)). Our second-degree benchmark \( \langle S, R \rangle \) therefore subscribes to cognitivism by constraining the stages of computations, whatever they may be, to process the same \( F \). Consequently, multiple realizability strikes again, this time with the potentially infinite number of algorithms [3,72,73] that can process the same stimulus features to produce the same behaviors. This segregates into the third degree of equivalence, where we constrain the hierarchical computations of \( A_{\text{DNN}} = a^1(…a^2(a^3(S))) \) to equate to those of \( A_{\text{Brain}} \) – in other words, similar hierarchical computations between generative stimulus features (\( F^1_{\text{Stim}}, F^2_{\text{Stim}}, …, F^L_{\text{Stim}} \)), DNN computations (\( F^1_{\text{DNN}}, F^2_{\text{DNN}}, …, F^L_{\text{DNN}} \)) and brain computations (\( F^1_{\text{Brain}}, F^2_{\text{Brain}}, …, F^L_{\text{Brain}} \)). Strictly speaking, the sequential order of DNN and brain computations should invert that of the generative model – in other words \( F^1_{\text{Brain}} = F^L_{\text{Stim}} \). The benchmark to evaluate the third degree of algorithmic equivalence becomes \( \langle S, R \rangle_{\text{F1}} \). Note that embedding the pivotal second degree of stimulus feature similarity into the third degree of algorithmic similarity clarifies the point made earlier – that we can only meaningfully compare brain and DNN computations (i.e., \( F_{\text{Stim}} = F_{\text{Brain}} = F_{\text{DNN}} \) for stage \( i = 1 … L \)) when we know that they categorize using the same stimulus features \( F \). Otherwise, we could literally compare the processing of apples and oranges across systems (that would similarly respond ‘fruit basket’ to the same stimuli). Thus, the argument of multiple realizability and the solutions developed (by using generative psychophysics) seamlessly apply from behavior to any other brain/DNN response measured at any level of granularity.

Third-degree equivalence of behavior because equivalent algorithms process equivalent \( F \)

At the third degree, we seek equivalence of the algorithms that process the same \( F \) from the same \( S \) to produce the same \( R \). We evaluate algorithmic similarity with the \( \langle S, R \rangle_{\text{F1}} \) benchmark and...
Trends in Cognitive Sciences

ExtractFace(image) = $a^3(\ldots a^1(image))_F$

ExtractTranslation(image) = $a^5(\ldots a^1(image))_F$

ExtractRotation(image) = $a^6(\ldots a^1(image))_F$

ExtractIdentity(image) = $a^7(\ldots a^1(image))_F$

$A_{DNN}(image) ::= \text{IdentifyFace(image)}_F ::= \ [
  \text{ExtractIdentity(}
  \text{ExtractRotation(}
    \text{ExtractTranslation(}
      \text{ExtractFace(image))}_F
    \text{))}_F
  \text{]}

(See figure legend at the bottom of the next page.)
seek to equate the outputs of \( L \) embedded computational stages \((F^1, F^2, \ldots, F^L)\) between the feedforward \( A_{\text{DNN}} \) and \( A_{\text{Brain}} \). Such algorithmic equivalence is like reaching for Neptune – it is beyond the current development of science. We realistically aim for an "algorithmic lift-off" with judicious use of generative psychophysics, as illustrated in Figure 2.

Suppose \( a^1(S) \) nonlinearly filters its input with banks of multi-scale, oriented Gabor kernels covering the input image. Subsequently, \( a^2(a^1(S)) \) max-pools the outputs of \( a^1(S) \) to reduce their dimensionality and implement their local invariance to translation. Following learning, \( a^2(a^1(S)) \) in DNNs would be similar to \( a^2(a^1(S)) \) models of visual cortex (V1/V2) cell responses in the primate brain [76]. However, instead of generically representing the entire image, a constrained \( a^2(\ldots (a^1(S))\ldots) \) could now deliver a compact representation of the task-relevant stimulus features \( F \). We can address the challenge of understanding how \( A_{\text{Brain}} \) and \( A_{\text{DNNs}} \) ultimately yield the \( F \) that categorize the stimuli while addressing the well-documented human higher-level invariances to other stimulus features. For example, we know that \( A_{\text{Brain}} \) identifies faces across parametric variations of many generated features (e.g., in Figure 2, background, illumination, translation, rotation, and the categorical feature of identity, as well as other features such as size and occlusion and categorical features of expression, aging, and so forth, that are not illustrated in Figure 2 [77]). Conversely, generalization performance decreases when we move away from the typical illumination from above, as well as for upside-down faces [78] and specific occlusions [79]. Relatedly, visual illusions (e.g., a 3D concave face moving from left to right produces a systematic illusory motion in the opposite direction [80]) can reveal priors (e.g., for convex surfaces) and logical dependencies of computations (e.g., \( A_{\text{Brain}} \) reconstructs face surfaces before identifying them).

To "lift-off" an algorithmic understanding we could apply generative psychophysics to feedforward \( A_{\text{DNNs}} \), as Figure 2 illustrates, and locate the sufficient sub-nested computations – \( a^1(\ldots (S))_i \) of \( A_{\text{DNNs}} \) – that generalize each stimulus feature in the same way that \( A_{\text{Brain}} \) does. This approach could map out the algorithmic functions that different sub-nested computations achieve, their relative complexity (e.g., four layers to generalize the face over backgrounds in Figure 2, but \( 4 + 3 \) to identify it), and their sequential dependencies (e.g., face extraction at layer 4, invariances to translation at \( 4 + 2 \), and finally identification at \( 4 + 2 + 1 \)). We could then ask whether a similar sequence of algorithmic functions organizes \( A_{\text{Brain}} \) across the hierarchically activated regions of the ventral and dorsal pathways.

Current evidence falls short of such algorithmic lift-off because of reliance on under-constrained first-degree comparisons of brain and DNN representations of the same 2D images. That is, when claims are made of equivalent brain and DNN layer responses [40,81,82], we rarely know the causal origin of the \( F \) being processed and represented in the brain (a second-degree question with more constrained benchmark \( \langle S, R_f; cf \rangle \) [83]), let alone the specific computations that cause them (a third-degree question, with even more constrained \( \langle S, R_{\text{brain}} \rangle \)). Thus, first-degree equivalence between brain and DNN activations exchanges one black box (the DNN) for another (the brain) [2,84]), teaching us little about either.

Figure 2. Third-degree algorithmic lift-off using the \( \langle S, R_{\text{brain}} \rangle \) benchmark. With our generative psychophysics we can locate the DNN layer that responds invariantly to different generative features of the stimulus [73]. Following training, a feedforward architecture (ResNet 7) responds invariantly to the background stimulus feature in layer \( L^4 \). Before \( L^4 \), layers \( L^{1-3} \) explicitly represent the background features. Invariance to face background suggests that the embedded algorithmic functions \( a^1(\ldots (S))_i \) of \( A_{\text{brain}} \) realize \( \text{ExtractFace} \) from the input images. A similar argument applies to the translation feature at \( L^8 \) and to the rotation feature at \( L^7 \). \( A_{\text{brain}} \) can be written in pseudo-language as \( \text{IdentifyFace(image)} \), itself defined as the sequence of embedded functions revealed by the generative psychophysics. \( A_{\text{brain}} \) therefore predicts \( \langle S, R_f \rangle \), by processing the \( F \) in a constrained sequence of algorithmic computations in different layers. In turn, we can turn to the hierarchically organized brain and ask whether \( A_{\text{brain}} \) computes these same \( F \) across its layers in the same algorithmic order as \( A_{\text{brain}} \). Abbreviations: a, embedded algorithmic function; A, algorithmic function; DNN, deep neural network; F, stimulus features; ID, Identity; L, layer; R, response; ResNet, residual network; S, stimulus/stimuli.
To reveal dynamic computations in the brain, a methodology was recently proposed based on generative psychophysics (albeit with simple features) [36]. This approach incorporates a well-defined algorithmic constraint into the task [86] to document the corresponding four main algorithmic stages of dynamic linear and nonlinear computations in different layers of brain networks that together transform stimuli into behavior. This approach in principle extends to the generative features of face, object, and scene categories.

The third degree will also need to address the classic argument that algorithms depend in part on the hardware that implements them [6,87], and how this hardware develops together with its software [88]. DNN architectures could embed well-documented properties of neuroanatomy and neurophysiology – for example, two retinas project in specific ways in two occipital cortices and then onwards into the ventral and dorsal categorization pathways, and their developmental trajectories (e.g., the initial processing of low-frequency visual information [89]). A step change from the typical feedforward architectures would enable recurrence to propagate top-down predictions that reverse the bottom-up flow [90]. In turn, predictions could guide active sampling of the visual input in the task [91] through sequences of "retinal" fixations on a generated 3D world to dynamically extract categorization features [34,35]. Lastly, additional constraints such as parsimony could be used to narrow down the search for appropriate algorithms [32]. Although third-degree algorithmic equivalence remains a tall order, DNN modeling more strongly constrained with a well-established methodology and evidence from vision and neuroscience would likely

**Box 3. Is the second degree really necessary?**

Several arguments can be developed against the need for our second-degree $\langle S, R_f \rangle$ to develop $A_{\text{brain}}$ as better models of $A_{\text{brain}}$. We argue that all imply $\langle S, R_f \rangle$, at least implicitly.

1. The ‘suitably rich’ $S$. A first-degree DNN modeler of $\langle S, R \rangle$ could object that the second-degree $\langle S, R_f \rangle$ remains unnecessary, as the argument would, go, boils down to the composition of a stimulus set $S$. A ‘suitably rich’ $S$ could dissociate responses $R$ across $A_{\text{brain}}$ and $A_{\text{brain}}$ without requiring a priori hypotheses about the features $F$ that generated $S$. However, this argument rests on composing the ‘suitably rich’ $S$ that enables such response dissociations. Without an explicit formalism of what $F$ might be, this composition would likely be guided by the modeler’s hypotheses about the images that contain (or not) the dissociating $F$ – in other words a second-degree endeavor, where $F$ remains latent in the mind of the first-degree modeler.

2. The asymptotic $S$. One could address the above shortcoming using brute force, piling millions of images into $S$ until the targeted responses are eventually dissociated. Even though this approach might asymptotically work (although self-driving cars and adversarial testing illustrate practical difficulties), the severely restricted experimental time of a neuroimaged brain makes brute force impractical. With neuroimaging, gathering a ‘suitably rich’ set (with low cardinality) that dissociates the responses of the systems at the first degree requires explicit $F$ hypotheses at the second degree.

3. A continuum of first-degree $\langle S, R \rangle$, differences between $A_{\text{brain}}$ and $A_{\text{brain}}$. Because different generative models of the same stimuli make as many different predictions about the $F$ of natural computations (e.g., from simple Gabors to multivariate 3D-shaped $F$), first-degree $\langle S, R \rangle$, performance comparisons between $A_{\text{brain}}$ and its $A_{\text{brain}}$ models would likely reveal a continuum of differences. Such a first-degree continuum, the argument would go, is all we need. However, the second degree introduces the qualitative difference of causal explanations, whereby $\langle S, R_f \rangle$ adds the cognitivist link because processing of $F$ to the first-degree $\langle S, R \rangle$ predictions.

A system of axes clarifies causal second-degree explanations (illustrated in Figure 1 in main text). Different sets of hypothesized generative features (e.g., Gabors vs. 3D multivariate shapes) each provide a different system of axes with explicitly defined dimensions (ideally, as few as possible) and units for each image that comprises $S$. Notions such as ‘suitable richness’ of $S$ and asymptotic $S$ all become defined within a generative space, with defined sampling density and coverage [74]. Performance overlap (or lack thereof) between $A_{\text{brain}}$ and $A_{\text{brain}}$ on any stimulus set $S$ becomes referenced to the explored regions of the space. Consequently, we now navigate along a defined system of axes to test generalization gradients, and we can target specific regions of uncertain performance overlap and increase sampling density to resolve the uncertainty.

To conclude, second-degree $\langle S, R_f \rangle$ is necessary to transform first-degree $\langle S, R \rangle$ from an odyssey in a randomly accessible territory into a plannable exploration that delivers systematically sampled explanatory maps of performance overlap with respect to (ideally understandable) generative features.
Concluding remarks

We have addressed the algorithmic equivalence between the categorizing brain and its DNN models in the context of categorizations of conspecifics and their surroundings. We discussed three embedded degrees that progressively increase the constraints that bear on the evaluation of algorithmic similarity: the first-degree equivalence of their categorization responses \( R \) to the same stimuli \( R_{\text{DNN}} \equiv R_{\text{Human}} \), the second-degree equivalence of the stimulus features \( F \) that produce these same behaviors \( F_{\text{DNN}} \equiv F_{\text{Human}} \), and the third degree of the L stages of algorithmic computations, reflected as layered brain features \( f_1, f_2, ..., f_l \) and DNN features between \( A_{\text{brain}} \) and \( A_{\text{DNN}} \). We have discussed how the first degree requires a modeling benchmark \( (S, R) \) that better incorporates the niche of the individual categorizer rather than the generic ground-truth responses of the DNN modeling community and the published convenience databases. At the second degree, we have emphasized the importance of developing a genuine psychophysics of human recognition and its DNN modeling that ‘breaks through the wall of pixels’ by developing 3D generative models of face, object, and scene categories. However, it is possible that the second degree may not be essential (Box 3). Nevertheless, such generative models will now be necessary to develop the type of algorithmic understanding that has been the hallmark of cognitive explanations and that is relevant for many cognitive questions and modeling approaches [74,93] (see Outstanding questions). Finally, we showed how generative psychophysics enables algorithmic lift-off by comparing the layered order of computations across regions of the categorizing brain and the layers of DNN models.

Acknowledgments

P.G.S. received support from the Wellcome Trust UK (senior investigator award 107802) and the Multidisciplinary University Research Initiative/Engineering and Physical Sciences Research Council (USA/UK; 172046-01).

Declaration of interests

The authors declare no conflicts of interest.

References

1. Dawson, M.R.W. (2013) Mind, Body, World: Foundations of Cognitive Science, Athabasca University Press
2. Kay, K.N. (2018) Principles for models of neural information processing, NeuroImage 180, 101–109
3. Polychron, Z.W. (1988) Computation and cognition: issues in the foundations of cognitive science, Behav. Brain Sci. 3, 111–112
4. Schyns, P.G. et al. (2009) Information processing algorithms in the brain, Trends Cogn. Sci. 13, 20–26
5. Cichy, R.M. and Kaiser, D. (2019) Deep neural networks as scientific models, Trends Cogn. Sci. 23, 305–317
6. Rumelhart, D.E. et al. (1986) Parallel Distributed Processing: Explorations in the Microstructure of Cognition (Vol. 1), Foundations, MIT Press
7. Kriegeskorte, N. (2015) Deep neural networks: a new framework for modeling biological vision and brain information processing, Annu. Rev. Vis. Sci. 1, 417–446
8. Lindsay, G.W. (2021) Convolutional neural networks as a model of the visual system: past, present, and future, J. Cogn. Neurosci. 33, 2017–2031
9. Zipser, D. and Andersen, R.A. (1988) A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons. Nature 331, 679–684
10. Franz, A. et al. (2013) DeViSE: a deep visual-semantic embedding model. In Advances in Neural Information Processing Systems (26), pp. 2193–2201, Curran Associates
11. Quest, O. and Martin, A.E. (2021) On logical inference over brains, behaviour, and artificial neural networks. PsyArXiv

Outstanding questions

How do published image datasets compare with naturalistic human visual sampling in a categorization niche?

How do DNNs predict the diversity of human categorization behaviors?

Can we train DNNs to indirectly engineer a generative (causal) model of 3D visual categories (e.g., conspecifics and their surroundings) from 2D image databases, or will we be restricted to directly engineering 3D graphics models that represent the real-world variations of one category at a time (e.g., face, or car, or city)?

Can DNN generalization gradients approximate human generalization gradients (e.g., to changes of illumination, scale, orientation, and categorization-relevant \( F_l \)?)

Can we develop different models of a given stimulus category to compare the processing of their generative features in the brain and DNN models?

Can DNN models account for classic visual illusions which reveal crucial differences between veridically versus humanly represented physical reality?

How deeply can DNNs model human visual cognition into scene semantics (e.g., from the features of ‘human at position \( x, y, z \)’ to those of ‘student hiker rescued from a Scottish hill after 2 days’?), and how can we construct generative models that incorporate such semantic constraints?

What algorithmic computations and representations can we extract from different modalities and granularities of brain measures?

How can we relate algorithmic computations and representations across modalities and granularities of brain measures?
21. Schyns, P.G. et al. (1998) The development of features in object concepts. Behav. Brain Sci. 21, 1–17.
22. Duan, Y. et al. (2022) Network mechanisms of dynamic feature selection for flexible visual categorizations. ArXiv:Published online May 9, 2022. http://dx.doi.org/10.48550/arXiv.2205.04300
23. Tanaka, J.W. and Taylor, M. (1991) Object categories and expertise: is the basic level in the eye of the beholder? Cogn. Psychol. 23, 457–482.
24. Mehler, J. et al. (2021) An ecologically motivated image dataset for deep learning yields better models of human vision. Proc. Natl. Acad. Sci. U. S. A. 118, e2011417118
25. Northcutt, C.G. et al. (2021) Pervasive label errors in test sets destabilize machine learning benchmarks. ArXiv:Published online March 28, 2021. http://dx.doi.org/10.48550/arXiv.2103.17479
26. Rajalingham, R. et al. (2018) Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. J. Neurosci. 38, 7255–7269.
27. Henrich, J. et al. (2013) The weirdest people in the world? Behav. Brain Sci. 33, 61–83.
28. Braunlich, K. and Love, B.C. (2019) Occipitotemporal representations reflect individual differences in conceptual knowledge. J. Exp. Psychol. Gen. 148, 1192–1203.
29. Jack, R.E. et al. (2018) Data-driven methods to diversify knowledge of human psychology. Trends Cogn. Sci. 22, 1–5.
30. Zhan, J. et al. (2021) Visual preferences reveals that face beauty is not universally perceived across cultures. Curr. Biol. 31, 2243–2252.
31. Jack, R.E. et al. (2012) Facial expressions of emotion are not culturally universal. Proc. Natl. Acad. Sci. 109, 7241–7244.
32. Ince, R.A.A. et al. (2022) Within-participant statistics for cognitive science. Trends Cogn. Sci. 26, 620–630.
33. Zhan, J. et al. (2019) Modelling face memory reveals task-generalizable representations across species. Nat. Hum. Behav. 3, 817–826.
34. Malcolm, G.L. et al. (2014) Beyond gist: strategic and incremental information accumulation for scene categorization. Psychol. Sci. 25, 1087–1097.
35. Henderson, J.M. and Hayes, T.R. (2017) Meaning-based guidance of attention in scenes as revealed by meaning maps. Nat. Hum. Behav. 1, 743–747.
36. Gauthier, I. et al. (1999) Activation of the middle fusiform ‘face area’ increases with expertise in recognizing novel objects. Nat. Neurosci. 3, 573.
37. Schyns, P.G. (1991) A modular neural network model of concept representation reveals hierarchical correspondence. J. Exp. Psychol. Hum. Percept. Perform. 17, 1000–1009.
38. Henderson, J.M. and Hayes, T.R. (2017) Meaning-based representation in a deep neural network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2921–2929. IEEE.
39. Koehler, K. et al. (2014) What do saliency models predict? J. Vis. 14, 14.
40. Linsley, D. et al. (2017) Ockloo: a web-based game for exploring the atoms of object recognition. ArXiv:Published online January 10, 2017. http://dx.doi.org/10.48550/arXiv.1701.02704
41. Murray, R.F. (2011) Classification images: a review. J. Vis. 11, 2.
42. Langlois, T. et al. (2021) Passive attention in artificial neural networks predicts human visual selectivity. In Advances in Neural Information Processing Systems (34), pp. 27084–27106. Curran Associates.
43. Daube, C. et al. (2019) Quantitatively comparing predictive models with the partial information decomposition. In 2019 Conference on Cognitive Computational Neuroscience, pp. 838–840. CCN.
44. Ince, R.A.A. (2017) Measuring multivariate redundant information with pointwise common change in surprise. Entropy 19, 318.
45. Goodfellow, I.J. et al. (2015) Explaining and harnessing adversarial examples. ArXiv:Published online March 20, 2015. http://dx.doi.org/10.48550/arXiv.1412.6572
46. Jacobsen, J.-H. et al. (2020) Excessive invariance causes adversarial vulnerability. ArXiv:Published online July 12, 2020. http://dx.doi.org/10.48550/arXiv.2007.04094
47. Gan, C. et al. (2021) ThreeDWorld: a platform for interactive multimodal physical simulation. ArXiv:Published online December 28, 2021. http://dx.doi.org/10.48550/arXiv.2007.04954
48. Yildirim, I. et al. (2020) Efficient inverse graphics in biological face processing. Sci. Adv. 6, eaax5979.
49. Jozwik, K.M. et al. (2021) Face dissimilarity judgments are predicted by representational distance in morphable and image-computable models. Proc. Natl. Acad. Sci. 119, e2115047119.
50. Goetschalckx, L. et al. (2021) Generative adversarial networks unlock new methods for cognitive science. Trends Cogn. Sci. 25, 788–801.
51. Mirza, M. and Osindero, S. (2014) Conditional generative adversarial nets. ArXiv:Published online November 6, 2014. https://doi.org/10.48550/arXiv.1411.1784
52. Ghosh, P. et al. (2020) GfG: generative interpretable faces. ArXiv:Published online November 25, 2020. http://dx.doi.org/10.48550/arXiv.2009.00149
53. 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. 7496–7507. IEEE.
54. Prinz, A.A. et al. (2004) Similar network activity from disparate circuit parameters. Nat. Neurosci. 7, 1345–1352.
55. Gutenkunst, R.N. et al. (2007) Universally sloppy parameter sensitivities in systems biology models. PLoS Comput. Biol. 3, e189.
56. Peterson, J.C. et al. (2021) Using large-scale experiments and machine learning to discover theories of human decision-making. Science 372, 1209–1214.
57. Xu, T. et al. (2018) Using psychophysical methods to understand mechanisms of face identification in a deep neural network. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 2057–2057. IEEE.
58. Dapello, J. et al. (2020) Simulating a primary visual cortex at the front of CNNs improves robustness to image perturbations. In Trends in Cognitive Sciences, December 2022, Vol. 26, No. 121101
Advances in Neural Information Processing Systems (33), pp. 13073–13087, Curran Associates

77. Wallis, G. and Rolls, E.T. (1997) Invariant face and object recognition in the visual system. Prog. Neurobiol. 51, 167–194

78. Valentine, T. (1988) Upside-down faces: a review of the effect of inversion upon face recognition. Br. J. Psychol. 79, 471–491

79. Sinha, P. et al. (2006) Face recognition by humans: nineteen results all computer vision researchers should know about. Proc. IEEE 94, 1485–1502

80. Hill, H. and Johnston, A. (2007) The hollow-face illusion: object-specific knowledge, general assumptions or properties of the stimulus? Perception 36, 199–223

81. Yamins, D.L.K. et al. (2014) Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc. Natl. Acad. Sci. 111, 8619–8624

82. Güçlü, U. and van Gerven, M.A.J. (2015) Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. J. Neurosci. 35, 10005–10014

83. Dwivedi, K. et al. (2021) Unveiling functions of the visual cortex using task-specific deep neural networks. PLoS Comput. Biol. 17, e1009267

84. Jaworska, K. et al. (2022) Different computations over the same inputs produce selective behavior in algorithmic brain networks. eLife 11, e79601

85. Maheswaranathan, N. et al. (2019) Universality and individuality in neural dynamics across large populations of recurrent networks. In Advances in Neural Information Processing Systems (32), pp. 1–13, Curran Associates

86. Griffiths, T.L. et al. (2015) Rational use of cognitive resources: levels of analysis between the computational and the algorithmic. Top. Cogn. Sci. 7, 217–239

87. Parisi, G. et al. (2019) Continual lifelong learning with neural networks: A review. Neural Netw. 113, 54–71

88. Jang, H. and Tong, F. (2021) Convolutional neural networks trained with a developmental sequence of blurry to clear images reveal core differences between face and object processing. J. Vis. 21, 6

89. van Bergen, R.S. and Kriegeskorte, N. (2020) Going in circles is the way forward: the role of recurrence in visual inference. Curr. Opin. Neurobiol. 65, 176–193

90. Thorat, S. et al. (2021) Category-orthogonal object features guide information processing in recurrent neural networks trained for object categorization. ArXiv Published online May 10, 2021. http://dx.doi.org/10.48550/2111.07988

91. Daube, C. et al. (2019) Simple acoustic features can explain phoneme-based predictions of cortical responses to speech. Curr. Biol. 29, 1924–1937

92. Daube, C. et al. (2019) Simple acoustic features can explain phoneme-based predictions of cortical responses to speech. Curr. Biol. 29, 1924–1937

93. Dayan, P. and Daw, N.D. (2008) Decision theory, reinforcement learning, and the brain. Cogn. Affect. Behav. Neurosci. 8, 429–453