Neuroscientists have long characterized the properties and functions of the nervous system, and are increasingly succeeding in answering how brains perform the tasks they do. But the question ‘why’ brains work the way they do is asked less often. The new ability to optimize artificial neural networks (ANNs) for performance on human-like tasks now enables us to approach these ‘why’ questions by asking when the properties of networks optimized for a given task mirror the given the available information [3], providing explanations for many observed visual phenomena.

ANNs are simulated networks of neuron-like units that are optimized by extensive training on a particular task through gradual adjustment of the connection strengths between units (Figure 1). These networks thus enable us to test the hypothesis that a particular mental or neural phenomenon observed in humans results from optimization for a specific task, by asking first whether that phenomenon arises spontaneously in a network trained on that task, and then, crucially, whether it does not arise when the network is optimized for other tasks. Thus, much as evolutionary theory explains the shape of a beak or length of a neck as not simply arbitrary species characteristics, but solutions to specific biological problems optimized by natural selection, we can explain specific characteristics of mind and brain as optimized solutions for specific computational problems faced by organisms. In the case of minds and brains, though, the optimization can occur either through evolution or through learning during development, or (more often) a complex combination of the two. Both forms of optimization offer possible answers to why minds and brains work the particular ways they do.

The idea that the particular problems the brain must solve strongly influence the computations it conducts is not new. For instance, David Marr noted long ago that ‘the nature of the computations that underlie perception depends more upon the computational problems that have to be solved than on the particular hardware in which their solutions are implemented’ [2]. This idea is reflected in the concept of an ideal observer that performs optimally on a perceptual task and whose underlying mechanisms, to a deeply theoretical enterprise of asking (and sometimes even answering) why they work the way they do. Among the next goals for research in this area would be to discover the underlying principles that explain why each optimized ANN produces the particular human-like phenomenon it generates (the ‘why of the why’).

Highlights

Understanding the mind and brain requires determining not only how they work, but why they work the way they do.

We argue that artificial neural networks (ANNs) provide a new method for addressing ‘why’ questions about the brain.

If an ANN optimized for a given task spontaneously produces a particular phenomenon previously observed in humans, but optimization for other tasks does not, that suggests that the phenomenon may result from optimization of the brain for that same task.

We review phenomena in vision and audition (e.g., specific illusions), and of cortical organization (e.g., specializations for face recognition) that arise spontaneously in ANNs optimized for specific tasks, providing possible explanations for why these phenomena occur in humans.

Acknowledgments

This work was supported by the National Institute of Health (R01 EY026605, U01 NS117071) and the Heidelberger Stiftung (to N.K.) and the German Research Foundation (He1583/7-1 to N.K., KL 946/7-1 to K.D.)

References


Correspondence:
katharina.dobs@psychol.uni-giessen.de
(K. Dobs)
in real brains. However, the ideal observer approach has proved less tractable for higher-level perceptual and cognitive processes. Here ANNs can help by discovering optimized (if not optimal) solutions to complex real-world computational problems [4]. When ANNs optimized for human-like tasks produce human-like phenomena, provides a possible explanation of why brains exhibit those phenomena, as well as an illustration of the very hardware independence Marr proposed.

One critical test of a causal explanation, the answer to a ‘why?’ question, is an intervention that removes the putative cause and asks if the effect still occurs [5]. A classic problem with evolutionary explanations of current species characteristics is that scientists cannot perform interventions that alter past objective functions and/or environmental constraints and rerun evolution (except in organisms with a very short generation time, like bacteria [6]). ANNs offer a solution to this problem for the case of optimization of minds and brains [7].

In this review we illustrate this general approach with examples of ANNs optimized for particular tasks that spontaneously produce known properties of minds and brains, thus explaining those properties as the possible results of optimizations for those tasks. We consider a broad definition of ANNs where an ANN is any network composed of simple computation units that loosely mimic real neurons, and where the connections weights between units are optimized according to an objective. The objective is again broadly construed, for example, the objective could be defined with reference to explicit supervision signals such as human-defined labels (as in supervised learning) or proxy tasks (as in self-supervised learning), or the objective could be to approximate the data distribution and capture the underlying structure of the data (as in generative modeling or
unsupervised learning). Although we focus on the role of task optimization, we note that ANNs also differ in other respects, including architecture, training data, and learning algorithms (Figure 1). These factors may interact with task constraints or may independently explain why or how a particular observed characteristic can arise.

Deep convolutional neural networks (CNN) as models for visual object recognition

In the field of computational vision, hand-engineered models of visual object recognition have been proposed for decades, but their performance lagged behind human abilities [8–10]. Then, in 2012, a deep CNN trained end-to-end on millions of labeled images burst on the scene with a performance close to that of humans on classification of real-world images [11]. CNNs thus offered the first plausible and image-computable models of how object recognition might work in the brain. Further, comparisons between such CNNs and primates showed a remarkable (though imperfect) match in their fine-grained behavioral performance [12] and in their internal activations [13–15]. Importantly, these networks were not trained to model primate object recognition, but only to classify images, making their fit to brains non-obvious and important. The (partial) fit tells us that these CNNs capture something about how vision works in the brain. It further suggests that merely optimizing for the same task can lead to similar solutions in brains and machines, despite the radical differences in their hardware and learning rules, supporting Marr’s conjecture that the brain’s visual algorithms are fundamentally shaped by the problems they are optimized to solve. This match between CNNs and brains thus also tells us something about why primate vision works the way it does: this is simply what an optimized solution to visual classification looks like! The success of CNN models of vision has inspired similar efforts in other domains, such as auditory perception [16,17] and language. Transformer-based large language models optimized for predicting the next word fit behavioral and neural data in humans, and the better the model performs on next word prediction the more closely it matches human data, suggesting that prediction may be part of what the human language system has been optimized for [18,19]. We next illustrate the power of ANNs to answer specific ‘why’ questions in cognitive science and neuroscience, with recent examples, starting with psychophysics.

Answering why questions about behavior

For over 150 years, psychophysicists have labored to characterize in detail the behavioral characteristics of human perceptual performance, documenting perceptual illusions and measuring precisely how visual acuity declines with stimulus eccentricity, how pitch discrimination is affected by the particular harmonics present in a tone, and how face recognition is affected by stimulus inversion. Testing whether ANNs optimized for certain tasks show similar phenomena (Figure 2) enables us to ask why humans exhibit these particular properties.

Audition

A pioneering study illustrating this strategy asked why human pitch perception exhibits the many well-established psychophysical characteristics it does [20]. To answer this question, the authors trained ANNs to estimate the fundamental frequency (F0), which is the perceptual correlate of pitch, from natural sound stimuli. The key finding was that ANNs that performed best at the task showed many of the classical characteristics of human pitch perception, such as changes in perceived F0 when stimuli were bandpass-filtered to control which harmonics were audible. Importantly, though, if the networks were instead trained on sounds without background noise, or sounds with unnatural spectra, these human-like behavioral signatures did not emerge. Similarly, if the input representation was altered so that it differed from that present in the auditory nerve, the model performed less well at pitch discrimination than humans. These and other findings thus explain many characteristics of human pitch perception as the result of optimization for...
the specific problem of extraction of the fundamental frequency of real-world sounds given human cochlear input.

Another study in the same genre [21] asked why human sound localization exhibits the many specific psychophysical characteristics it does, such as frequency-dependent use of interaural time and level differences, localization dominance of sound onsets, and limitations on the ability to localize multiple concurrent sources. To find out, the authors built an ANN simulating the outer ears and head/torso with impulse responses recorded from a physical model of head and ears, and simulating a cochlea with a set of human-like bandpass filters. They then trained this model end-to-end to localize sounds generated in a virtual world with realistic background noise and reverberation. After training, they tested the model on a wide range of classic psychophysical tasks, and found that the model duplicated many previously established phenomena of human auditory localization, including those mentioned above. These findings suggest that many psychophysical properties of human auditory localization reflect optimizations for the specific problem of sound localization in natural environments given the fixed properties of the peripheral human auditory system.

Vision
Classic visual psychophysical phenomena are also starting to be explained as the result of optimizations for particular tasks. For example, many visual illusions arise spontaneously in networks optimized for visual tasks (Figure 3), and which illusions the CNNs replicate depends on the task each CNN is optimized to solve. Other psychophysical phenomena found in CNNs trained on object recognition (but not CNNs with random weights) include set size effects in visual search [22], a hallmark of human visual search performance, and human-like mirror confusion and scene incongruence effects (i.e., improved performance when an axe is embedded in a forest compared to a supermarket) [23], suggesting that these phenomena result from an optimization for object
recognition in natural scenes. Importantly, however, object-trained CNNs did not show several properties of the human visual system related to 3D processing, occlusions or invariance to surfaces, and part-based processing [23], suggesting that these phenomena might result from other tasks beyond object classification. While these results support the role of specific tasks for properties of human object perception, a few phenomena of human object perception, such as relative size encoding, emerge in randomly initialized CNNs even before any training [23]. This finding emphasizes the role of network architecture (Figure 1) – in addition to the training task – and shows how random feedforward connections can already give rise to useful features.

Human face recognition exhibits a number of distinctive and well-documented behavioral ‘signatures’ such as overall high accuracy that drops significantly when faces are unfamiliar, or presented upside down, or originate from an ethnicity/race the observer is less familiar with. Why does face perception exhibit these properties? A recent study from our group used CNNs to test the hypothesis that these signatures result from an optimization for the task of fine-grained face discrimination [24]. We found that many of these face-processing signatures are
found in CNNs trained on face recognition, but not in CNNs trained on object recognition, suggesting that these phenomena may arise from optimization specifically for face recognition [23, 25]. Alternatively, however, these properties could simply emerge from face-specific experience (without the need to discriminate the faces individually). To disentangle these two hypotheses, we manipulated the task and the amount of face-specific experience by training CNNs on face detection (categorizing all faces into one category) in addition to object recognition [24]. We found that when including the same amount of face experience, but without training the CNNs to perform fine-grained face discrimination, the classic human signatures of face perception were absent or weak, implicating optimization for face discrimination, rather than simply face experience, in these phenomena (Figure 2B).

Taken together, these findings demonstrate how varying the experience and task in CNNs enables us to go beyond documenting and reporting behavioral phenomena of the visual system, to asking why it exhibits these phenomena in the first place. Of course, simply saying that phenomenon X results from optimization for task Y begs the further question of why optimization for Y produces X, the ‘why of the why’, which can be further pursued in a variety of ways (Box 1).

**Answering why questions about brains**

Deep neural network models can inform ‘why’ questions not only about human behavior but also about the organization and function of the brain. Comparing ANNs optimized for certain tasks to minds and brains enables us to ask why the brain exhibits particular properties (Figure 4). We start with the earliest stage of the visual system and follow it downstream.

**The organization of early stages of visual processing**

Why do primates have much lower spatial resolution in the visual periphery than in the center of vision? This fact is usually explained in terms of the high metabolic and wiring cost of photoreceptors and their connections. But one study tested a different hypothesis by training a neural network on a visual search task entailing saccade-like translations of the input image over a receptor lattice [26]. The position and spatial resolution of each receptor was optimized over

---

**Box 1. The why of the why**

Comparing brains to ANNs can reveal the design constraints that may have shaped brains, answering ‘why’ questions about observed neural or behavioral phenomena. However, simply invoking specific objective functions and constraints as explanatory primitives begs another question.

What specific aspects of a particular task are critical for an optimized ANN to reproduce a neural/behavioral phenomenon of interest? Two examples illustrate initial strategies for delving deeper into the ‘why of the why’.

Evidence that specialization for face and object recognition in the brain results from optimization for both tasks comes from the finding that a similar segregation arises spontaneously in ANNs jointly trained on both tasks [44]. But this result begs the question of why face-processing segregates in both brains and networks (the ‘why of the why’). That is, what is it about the task of face recognition that requires separate machinery? One hypothesis is that functional segregation will arise only for categories in which the exemplars to be discriminated all share the same basic shape. Initial evidence for this hypothesis was found when car discrimination segregated from object classification in networks trained on both, but no segregation was found for discrimination of faces from two data sets with different low-level image properties. However, segregation of the highly heterogeneous visual category of food in both networks [44] and brains [72] suggests that this account is incomplete at best.

In answer to why human sound localization is dominated by the first part of a sound (the ‘precedence effect’), a recent study found that ANNs trained to localize naturalistic sounds show the same effect [21]. But why should this effect arise in both networks and brains? To test the hypothesis that the precedence effect represents a solution to the problem caused by echoes, which can reflect off surfaces far from the original sounds source, the authors trained a new ANN on sound localization using sounds generated in an anechoic environment. Consistent with their hypothesis, no precedence was observed in this network, suggesting that the effect represents a solution to the ambiguity caused by echoes.
training, and produced a higher-resolution ‘fovea’ at the center and a lower-resolution periphery. Importantly, when the model was allowed to make non-biological image transformations like zooming, the fovea-like organization did not arise, suggesting that the organization of the primate fovea may have evolved in part to enable efficient sampling of visual information over saccades.

Another study showed that CNNs trained on object recognition with a human blur profile at the input stage outperformed networks trained on input images with steeper or shallower blur profiles, or full resolution images [27]. Both findings suggest that a blurred visual periphery may be a feature, not a bug, reflecting evolutionary optimization for object recognition and/or efficient sampling of visual information across eye movements.

Moving along the visual pathway, another recent study asked why V1 has the functional characteristics it does, by building these properties into the front of a CNN (VOneNet) and training it on Imagenet [28]. The resulting network was more robust to adversarial attacks and common image corruptions than state-of-the-art networks. Further, each of the properties of V1 built into the
model contributed to the network’s robust performance, as removal of any one of them decreased robustness. These results suggest that many properties of V1 reflect evolutionary optimizations for robust image classification. Another study tested whether processing of color and luminance would spontaneously segregate in CNNs trained on object recognition, as they do in early visual cortex [29]. The authors trained multiple instances of Alexnet and found a high degree of segregation of chromatic and achromatic information across CNN instances. Moreover, the degree of segregation in a network was correlated with its performance. This finding suggests that the segregation of color and luminance processing in the human visual cortex may also result from optimization for real-world object recognition.

Another study attempted to understand the systematic spatial organization of early visual cortex using computational models built on self-organizing principles [30]. Specifically, by transforming the visual input space into a tuned 2D map, with each unit tuned to some aspect of the visual space and nearby units having similar tuning, a repeating map topography emerged, similar to the primate visual system. This suggests a role for biologically plausible self-organizing principles (which in turn reflect approximate solutions to wiring cost minimization) in shaping the organization of the early visual cortex (Figure 5A).

**Higher-level stages of visual cortex**

From retinotopic cortex, visual processing diverges into a ventral object recognition (or ‘what’) stream and a dorsal (‘where’) stream processing object location and visually guided action. One of the first studies of the genre highlighted in this review asked why the visual cortex is organized into these two streams [31]. To find out, the authors trained two versions of a simple three-layer connectionist network: one in which the nodes in the hidden layer were split between those connected only to the shape output nodes and others connected only to the location output nodes, and another version in which all hidden units were connected to all output nodes. The authors found that the split networks outperformed the unsplit network, but only when more hidden units were allocated to the (more difficult) ‘what’ task. A related connectionist study used a modular architecture in which the different modules compete to learn the task, resulting in a partitioning of the task into multiple functionally distinct subtasks, with a distinct module allocated to each [32]. This problem was later revisited with modern deep neural networks by training CNNs for two visual tasks simultaneously [33]. Specifically, the authors manipulated the relatedness of the two tasks to test the hypothesis that the need to perform two unrelated tasks (like the ‘what’ and ‘where’ task during object processing) results in the emergence of segregated processing streams dedicated to each task. In the CNN trained on related tasks, the majority of units contributed to both tasks, whereas in the CNN optimized for unrelated tasks, units often contributed disproportionately to a specific task, and the degree of specialization increased with progressive layers. Taken together, these and other findings [34] suggest that segregation of function in the visual system results from the optimization for multiple tasks with different computational goals (Figure 5A).

Other recent studies have asked why the ventral visual pathway is organized the way it is, with small regions selective for the specific categories of faces [e.g., occipital face area (OFA), fusiform face area (FFA)] [35], places [parahippocampal place area (PPA)] [36], bodies [extrastriate body area (EBA)] [37], and words [visual word form area (VWFA)] [38] embedded in larger cortical regions exhibiting gradients of weak preferences for mid-level features (e.g., [39–41]) (Figure 5B–D). One such study [42] trained a CNN on object and scene categorization and found that the network organized itself into units with a central image bias and units more selective for the background of images. This finding suggests that the dissociation between fovea-biased and periphery-biased regions of the ventral pathway [43], and the functional properties of each, may reflect an optimization for the classification of images from different visual categories (Figure 5A).
Figure 5. Aspects of cortical organization that have been informed by artificial neural network (ANN) models. (A) ANN models trained on basic perceptual tasks provide computational explanations for multiple aspects of observed cortical organization. ANNs mirror the functional organization of the early visual cortex [30] and hierarchical organization of auditory cortex [16,17], show layer-wise correspondences between neural responses in primary versus higher-level regions of visual cortex [13–15], and brain-like functional differences between fovea-biased and periphery-biased cortex in the ventral pathway [42]. Functional dissociations emerge in ANNs for faces versus objects and food versus objects [44], for the processing of visual words [45], for ‘what’ versus ‘where’ pathways in vision [31,33,34], and speech versus music in auditory cortex [16]. Several ANN models with connectivity constraints and topographic representations have further been used to account for the systematic spatial organization of the high-level visual cortex. These models primarily differ in what connectivity constraints are assumed and how the ANN representational space is mapped to the two-dimensional topographic space. (B) Localized face selectivity emerged in a topographic ANN model trained for

(Figure legend continued at the bottom of the next page.)
In another recent study, we asked why the visual cortex exhibits such a high degree of specialization, and why it does so for some categories but apparently not for others [44]. Using network lesioning methods (Figure 4), we found that a CNN jointly trained on both face and object recognition spontaneously segregated itself into separate systems for faces and objects, suggesting that human brains show this organization as a result of joint optimization for both tasks (Figure 5A and Box 1). In a related line of work, ANNs were harnessed to account for the emergence of visual word selectivity in the ventral visual cortex [45]. The authors simulated the late acquisition of reading abilities in humans by training an ANN model in two phases, first on general image recognition, and subsequently on both image and word recognition. A subset of single units exhibited strong word-selectivity following training, thereby capturing known properties of the VWFA (Figure 5A). This model provides a further demonstration of the computational utility of functional segregation, as well as supporting the cortical-recycling theory of development by demonstrating how a network optimized for generic object recognition may be repurposed to additionally perform visual word recognition by co-opting a small subset of units for this task.

Within the domain of face perception, a classic view holds that while face identity is processed by the ventral pathway (e.g., including OFA and FFA), facial expressions are processed in a lateral temporal pathway including the posterior superior temporal sulcus (STS). However, this view has been challenged by reports that facial expressions can also be decoded from ventral areas, and identity information can be decoded from lateral regions [47,48]. Recently, CNNs were used to inform this debate. If segregation of facial identity and expression processing is required from a computational point of view, then CNNs trained to recognize identities should outperform CNNs trained to recognize facial expressions on face identification and vice versa. Interestingly though, expression-selective units spontaneously emerge in CNNs trained for facial identity (and vice versa) [49,50], and these expression-selective units show human-like characteristics [51]. These results suggest that face identity and expression are processed interdependently, and that functional segregation of these two processes is not necessarily expected on computational grounds. Moreover, this finding shows how CNNs can be used not only to ask why some processes are functionally segregated, but also to explain why other processes are integrated.

Functional organization of auditory cortex
Can the functional organization of auditory cortex also be understood as resulting from task optimization? Recent studies have trained CNNs on auditory tasks and found correspondences between the trained models and human auditory cortex, with earlier stages predicting primary auditory regions and deep stages predicting nonprimary regions [16,17,52,53]. Further, models trained on multiple auditory tasks had the best overall predictivity for neural responses [17]. Training CNNs that branched at different layers on speech recognition and musical genre classification further revealed that the networks that performed best shared early processing stages across tasks, but engaged separate pathways for speech and music at later stages [16]. This branched network matched human task performance, exhibited a similar pattern of errors, and predicted voxel responses in human auditory cortex, despite being optimized only for task performance. This work suggests that the functional organization of human auditory cortex reflects optimizations for the human auditory tasks including speech and music recognition (Figure 5A).
Unsupervised training to ask ‘why’ questions about development

An important new direction in the use of ANNs to inform human behavioral and neural organization is the recent advent of unsupervised models trained not on labeled data but on suitable unsupervised proxies for labels that are obtained via simple image manipulations. The representations learned via these ‘semi-supervised’ learning models now compete with supervised models in their object recognition behavior [54]. Recent papers showed that ANNs trained with deep contrastive unsupervised methods can predict neural responses to images in the monkey [55] and human [56] ventral visual pathway as well as supervised models, and also exhibit more human-like error patterns [55] than supervised models. This finding suggests that the biologically unrealistic form of label-based feedback received by supervised models is not necessary to achieve human-like neural or perceptual phenomena, filling in the explanatory gap created by supervised models which rely on millions of semantic labels and are thus implausible as models of biological learning. All these unsupervised models sample different views (augmentations) of the same image and are trained to maximize agreement between their representations; unlike supervised models, the labels for semi-supervised ANN models (i.e., the views) are accessible in humans, for example, through retinal distortions or saccades [56]. While unsupervised learning algorithms are thus clearly more biologically plausible than their supervised counterparts, both techniques still align in an overarching goal, for example by providing proxy labels to achieve object recognition. Importantly, these models can help us get closer to answering ‘why’ questions from the lens of postnatal development during which labels are rarely provided.

Strengthening evidence for optimization arguments

The research strategy described here entails inferring that when an ANN optimized for a given task spontaneously produces behavioral or neural characteristics previously described in humans, but optimization for other tasks does not produce the same characteristics, that suggests that these human characteristics reflect optimization for that task. It is important to note that no claims about mechanistic similarity between ANNs and brains are made in this framework since it is the optimization, and not the precise mechanism, that is driving the explanations. Different models could have the same input–output function but vary in their mechanistic plausibility; yet if they all exhibit the neural/behavioral phenomenon of interest, then this suggests an important link between optimization for the function and the emergent phenomenon.

As in any scientific domain, we can never definitively prove that an optimization hypothesis is true, and future evidence could always overturn it. However, the strength of the evidence that phenomenon X resulted from task optimization Y will increase with (i) the breadth of networks varying in hyperparameters but sharing the same task optimization Y that all produce phenomenon X, and crucially, (ii) the range of networks optimized for different tasks that do not produce X. We can further attempt to find the ‘minimal’ sufficient condition for phenomenon X to emerge in ANNs, that is, the smallest possible difference between ANNs that produce a cognitive/neural phenomenon and ANNs that do not. Moreover, it is important to determine whether the effect in the network is of similar magnitude to that observed in humans, such that its emergence in networks provides a compelling explanation of the human phenomenon. Finally, we can increase the rigor of the overall scientific program outlined here by reducing experimenter degrees of freedom with preregistrations of our specific optimization hypotheses and the ANNs we will use to test them.

Further, there may be a few cases in which we can experimentally test a proposed answer to a ‘why’ hypothesis. When the answer refers to optimization over development, it is sometimes possible to create in animals, or find in humans, conditions in which the relevant experience differs. In these cases, we can test whether developmental optimization that hinges on the availability of
particular experience can explain the emergence of a given phenomenon. For the case of optimization over evolution one can sometimes test hypothesized answers to ‘why’ questions that appeal to optimization by comparing across species with different niches [57].

Limitations of optimization arguments

While the research strategy described here is potentially powerful, it is subject to several important caveats. For one, if the characteristic in question emerges only for certain sets of network hyperparameters, the argument loses force, showing merely that the characteristic can emerge from the objective function in question, rather than that it is either likely or sure to. It is therefore essential that researchers do not cherry-pick among the large space of possible hyperparameters if they are trying to use ANNs to answer ‘why’ questions. Better yet, researchers should kick the tires on network hyperparameters (e.g., using different random initializations [58]) to test the robustness of the emergence of the particular phenomenon in question. Second, even when the characteristic in question does not emerge in control models (e.g., an ANN with random initialization or a different training objective), a thorough optimization of all relevant parameters for the control models is essential to make general statements about the failure of control conditions to yield the phenomenon in question [59].

Further, because multiple objective functions can in principle share a common solution, a model that replicates observed neural responses is not guaranteed to have the same objective function as the brain. By contrast, in some cases where multiple task optimizations produce the same characteristic [60], researchers can attempt to understand what those tasks have in common such that they produce that characteristic or what other emergent phenomena are consistently shared across these diverse ANNs [61], and researchers can then test those hypotheses by training networks with new objective functions. Indeed, this approach will ultimately enable us to explore the ‘why of the why’ by not just testing which task optimizations lead to a given solution, but what exactly it is about that task or training set that leads to that solution (Box 1). The converse challenge is that a given objective function may have multiple solutions (i.e., multiple local minima), so there is no guarantee that even the correct objective function will lead to a match to the brain. One proposal for addressing this problem is the ‘contravariance principle’ according to which the space of possible solutions is smaller for more complex tasks, suggesting that the approach advocated here may be most effective when applied to higher-level perceptual and cognitive processes [62]. In any case, the upshot of these critiques is that while ANNs are powerful tools that are opening up new avenues for addressing long-standing why questions in cognitive science and neuroscience (see Outstanding questions), they cannot do the job on their own, and we still need to think hard about the computational principles underlying the emergence of phenomena from network optimization.

Concluding remarks

We have argued here that much as evolution offers a framework for explaining why organisms have the characteristics they do, ANNs give us a method for asking why the human brain has the characteristics it does. Importantly, though, the solutions found in both evolution and ANNs are optimized but not optimal: evolution depends on prior conditions and a dynamically changing environment, and optimization in ANNs is always tested within a particular set of hyperparameters. Nonetheless, these two frameworks enable scientists to move beyond the mere collection of facts, and the exploration of underlying mechanisms, to approach some of the deepest theoretical questions about why organisms and minds work the way they do. And whereas evolutionary theories are sometimes criticized as ‘just so’ stories because of the difficulty of testing those theoretical accounts, with ANNs researchers can test their explanations by altering the objective function, learning rules, training data, or architecture. We have described here some exciting first steps
where this strategy has been successfully applied to a wide array of behavioral phenomena such as auditory pitch perception and sound localization, visual illusions, and higher-level visual processing of objects and faces. It has also illuminated multiple aspects of the functional organization of visual and auditory processing in the cortex. Opportunities abound for extending this work into new domains, including language [18,19,63], navigation [64–67], motor control [58,69], and higher-level cognition [70,71]. There is grandeur in this lens on cognition, which enables us to query the characteristics of our very own minds and brains with the quintessentially human question: why?

Acknowledgments
We thank Greta Tuckute for valuable comments on the manuscript, and members of the Kanwisher laboratory for fruitful discussions and feedback. This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation; project number 222641018–SFB/TRR 135), ‘The Adaptive Mind’, funded by the Excellence Program of the Hessian Ministry of Higher Education, Science, Research and Art to K.D., National Institutes of Health grant DP1HD091947 to N.K. and National Science Foundation Science and Technology Center for Brains, Minds, and Machines.

Declaration of interests
The authors declare no competing interests in relation to this work.

References