Psychonomic Bulletin and Review (2021) 28:454–475
https://doi.org/10.3758/s13423-020-01825-5

THEORETICAL REVIEW

Artificial cognition: How experimental psychology can help generate explainable artificial intelligence

J. Eric T. Taylor[1,2] · Graham W. Taylor[1,2]

Accepted: 2 October 2020 / Published online: 6 November 2020

Abstract
Artificial intelligence powered by deep neural networks has reached a level of complexity where it can be difficult or impossible to express how a model makes its decisions. This black-box problem is especially concerning when the model makes decisions with consequences for human well-being. In response, an emerging field called explainable artificial intelligence (XAI) aims to increase the interpretability, fairness, and transparency of machine learning. In this paper, we describe how cognitive psychologists can make contributions to XAI. The human mind is also a black box, and cognitive psychologists have over 150 years of experience modeling it through experimentation. We ought to translate the methods and rigor of cognitive psychology to the study of artificial black boxes in the service of explainability. We provide a review of XAI for psychologists, arguing that current methods possess a blind spot that can be complemented by the experimental cognitive tradition. We also provide a framework for research in XAI, highlight exemplary cases of experimentation within XAI inspired by psychological science, and provide a tutorial on experimenting with machines. We end by noting the advantages of an experimental approach and invite other psychologists to conduct research in this exciting new field.

Keywords Hypothesis testing · Comparative cognition

> In the history of science and technology, the engineering artifacts have almost always preceded the theoretical understanding
>
> - Yann LeCun

Introduction

Machine learning (ML) is changing modern living at a rapid pace. Medicine, finance, and transportation are among the many fields poised for transformation by the proliferation of machine learning models that can outperform their human counterparts. These algorithms make decisions that have significant consequences for the health and happiness of the people who use them—for example, identifying whether a blip on a scan has the potential to be cancerous, to apply the brakes in an autonomous vehicle, or to award a loan. One challenge facing the creators of these algorithms is that modern artificial intelligence (AI) solves problems in inscrutable ways–the so-called black-box problem. Because these models refine themselves autonomously and with an idiosyncrasy beyond the scope of human comprehension and computation, it is often impossible for a model's user or even creator to explain the model's decision.

Due to the transformative promise of AI at scale and the urgent lack of satisfying explanations for AI decision-making, there is increasing political, ethical, economical, and curiosity-driven theoretical pressure on ML researchers to solve the black-box problem, creating a sub-field called explainable artificial intelligence (XAI). In this paper, we advance an interdisciplinary approach to XAI known as Artificial Cognition (cf. Ritter et al., 2017), drawing heavily on the tradition of experimentation developed within cognitive psychology. This is a call for a new field.

In short, we draw parallels between the black-box problem in XAI and a similar epistemic challenge faced by cognitive psychologists. This paper is specifically written to engage cognitive psychologists in a current and applied challenge where their skill set should prove invaluable:

✉ J. Eric T. Taylor
j.eric.t.taylor@gmail.com

[1] The Vector Institute for Artificial Intelligence, Toronto, ON, Canada

[2] The School of Engineering, The University of Guelph, Guelph, ON, Canada

solving black-box problems. Although XAI is producing a large literature that attempts to explain machine behavior, we note that contributions from psychology have been scarce; we hope this paper will motivate other researchers to join us. To get started, we establish a case for cognitive psychology's contribution to XAI; we provide a review and taxonomy of the current state of XAI and offer suggestions for how psychologists can contribute; we define Artificial Cognition and propose guidelines for how to conduct experiments with machines and highlight excellent cases from the ML literature; and we provide a short tutorial on some critical differences between experimentation with humans and machines and the advantages of the Artificial Cognition approach.

The black-box problem in machine learning & psychology

An automotive engineer would have no difficulty explaining the function of an automobile were they able to look under the hood and see the components involved. The same cannot be said for a computer engineer and modern deep neural networks. After millions of training iterations, once the model has reached competence on a given task, the resultant set of parameters and operations that characterize the model would be so complex that no human could meaningfully say, from looking "under the hood", how it makes a decision[1]. Indeed, if we arbitrarily assume the reader has 50 years left to live, they would not be able to read a full list of parameters in some practical convolutional neural networks (CNNs) trained for image classification in the time they have left on Earth. This is what we mean by the *black-box problem*: We currently cannot meaningfully understand an AI's behavior as a function of its inner workings. Lillicrap and Kording (2019) expressed it eloquently:

> "After training we have the full set of weights and elementary operations. We can also compute and inspect any aspect of the representations formed by the trained network. In this sense we can have a complete description of the network and its computations. And yet, neither we, nor anyone we know feels that they grasp how processing in these networks truly works."

The above formulation of the black-box problem is also the best-case scenario, where the researcher could feasibly "look under the hood" of the AI in question. As has been noted, much of the source code for the models and agents we are interested in are proprietary and understandably secret, further complicating the challenge of XAI (Rahwan et al., 2019). In most cases, we *only* have access to input and output, and must infer the decision-making process. As researchers interested in explaining AI decisions, we have access to the input, the output, and a mass of hidden functions that are either inscrutable because we are not allowed to look at them, uninterpretable because we have no way to understand them, or both.

This formulation of the black box problem is the same epistemic challenge championed by cognitive psychologists for the last 150 years. Going back to at least 1868, psychologists have used behavioral experiments to infer the properties of invisible mental processes, as when Donders recorded response times and used the subtractive method to identify stages of perceptual processing and response selection (Donders, 1868). In the intervening century-and-a-half, cognitive psychology has yielded robust models of perception, attention, memory, language, and decision-making, all without ever observing these processes directly. The tradition of inferring cognitive models of human behavior from experimental data is a different version of the same black-box problem faced by XAI.

We advance an alternative and complementary approach to XAI inspired by the success of cognitive psychology and cognitive neuroscience. Our proposal is to describe black-box processing in AI with experimental methods from psychology, the same way we do it with humans. Whereas computer scientists explain artificial intelligence by tinkering under the hood of the black box, cognitive psychologists have developed a science of behavior that works without opening its black box. Using the experimental method, cognitive psychology employs carefully crafted stimuli (input) and measures the corresponding behavior (output) to make causal inferences about the structure and function of the human mind. We should apply the same approach to artificial minds. Instead of altering AI architecture or generating post hoc explanations for how AI reaches its decisions, we can develop satisfying models of mind without interfering with the AI's black box. In addition to providing models for artificial minds with the goal of explaining their decisions and processes, an experimental approach to XAI could provide guiding insights for advancing new design.

Ethical and political need for XAI

Why should we care how AI makes decisions, so long as it makes good decisions? Deep learning algorithms are wired to detect even the faintest meaningful correlations in a data set, so decisions based on a trained model's parameters ought to reflect correlations in reality. The challenge is that the parameters represent truth in a training data set, rather than truth in the world. There are numerous vulnerabilities to users that arise from this discrepancy. For example, data sets that are biased against a particular group will yield predictions that are biased against that particular

[1] There are some notable examples of analyzing small neural networks by hand in the days before "deep learning" (cf. Hinton, 1986).

group (Barocas et al., 2019); commercial models that are trained on biased data sets will treat underrepresented groups unfairly and inaccurately (Buolamwini & Gebru, 2018); vehicles that are trained in test circuits may not generalize adequately to real roads (Kalra & Paddock, 2016); and machines that are vulnerable to adversarial attacks may backfire or behave dangerously (Gu et al., 2019). It behooves us as creators, vendors, and users of these technologies to understand how and why things can go wrong.

The black-box problem in AI has recently motivated legislative bodies to give their citizens a right to know how AI makes its decisions. In 2018, the first such legislation came into effect in the EU, a major market for AI products (of the European Union, 2016; Goodman & Flaxman, 2017), with more stringent versions in other jurisdictions. These laws give citizens a right to understand how AI makes decisions when it makes decisions on their behalf. France's take on right-to-know legislation includes the communication of parameters, weights, and operations performed by the AI in some circumstances (Digital Republic Act, law no. 2016-1321). Policy analysts currently debate to what extent these laws are legal or even helpful (Edwards & Veale, 2018). XAI is a case where an ethical need gave rise to a political promise that has become a legal quagmire.

We believe the mandated communication of black-box parameters and operations would be meaningless to any user, harmful to the scientific and entrepreneurial potential of developers, and completely beside the point of delivering explanations of behavior. The real legal and ethical challenge for XAI is to reveal explanations that users find satisfactory and that make accurate and reliable predictions about AI behavior in general and fairness and safety in particular.

Satisfaction is an important consideration because it determines trust and use. Human trust in machine behavior depends on many factors, including whether the task is deemed formulaic (Castelo et al., 2019), and whether the agent is anthropomorphic (de Visser et al., 2016), not whether the math is transparent. For some people, trust will be determined by how the machine behaves in sensitive situations. In these cases, a satisfactory explanation must be able to balance different outcomes. For example, we might expect that when users are endangered, the AI should evaluate the best outcome to minimize harm. Satisfactory explanation in the case of car accidents resulting in harm are complicated by the fact that people have different ideas of how AI should act in ethical quandaries depending on their culture (Awad et al., 2018), or whether they've been primed to take the perspective of passenger or pedestrian (Frank et al., 2019). In other words, trust and ethics are both flexibly interpreted, and explanations will only be satisfactory if they allow a user to judge whether the decision was appropriate in that situation. To this end, we argue that explanations must be causal, rather than correlative (Pearl, 2019). We contend that, as in human psychology, correlation is insufficient to explain behavior in machine psychology. It is not enough to say "we think your loan was denied by the bank's AI because your profile is low on variables x, y, and z, but we can't be sure." We ought to have a causal model that survives experimental attempts at falsification.

Machine behavior as XAI

The researchers who explain AI behavior have historically been the same researchers who invent, develop, and train AI. A large sector of XAI research (reviewed more fully in "A review of XAI for psychologists") aims to generate explanations through more interpretable architectures, or by introducing a second AI that examines another AI's decision and learns to generate explanations for it. In other words, a popular approach to explainability in AI is more AI. Some XAI researchers are beginning to explore alternate paths to explanation, as evidenced by recent calls for the primacy of empirical study.

In 2017, Tim Miller published an influential call to improve the quality of explanations in XAI by drawing on the vast literature of philosophy, psychology, and cognitive science (Miller, 2017a). In this review, Miller outlined what makes a good explanation and established the need for validation of XAI models with human data. This review was followed by a survey that assessed the degree to which this nascent literature is integrated with the social sciences. The researchers asked whether the references of this small corpus dealt with explanations that humans found satisfying, whether the explanation data was validated with humans, and whether the referred articles were published in journals or conference proceedings outside computer science (Miller et al., 2017b). As of 2017, XAI and the social sciences were largely disconnected, establishing a case for an interdisciplinary approach. Around the same time, popular press articles began to capture public attention in the need for human-centric XAI (Rahwan & Cebrian, 2018; Kuang & Can, 2017), and questions about explainability became more common in media and interviews with popular figures in the machine learning world. Consider this quotation from a 2017 interview with Peter Norvig, a Director of Research at Google, presaging the proposal for machine behavior as an alternative to AIs that try to explain themselves (emphasis ours; Norvig, 2017):

> What cognitive psychologists have discovered is that when you ask a human [how they made a decision] you're not really getting at the decision process. They

> make a decision first and then you ask and then they generate an explanation and it may not be the true explanation. So we might end up being in the same place with machine learning systems where we retrain one system to get an answer and then we train another system to say 'given the input of this first system now it's your job to generate an explanation'. We certainly have other ways to probe. Because we have the system available to us we could say 'well what if the input was a little bit different—would the output be different or would it be the same?' So in that sense there's lots of things that we can probe. And I think having explanations will be good. We need better approaches to that—**it should be its own field of study**.

Recently, an interdisciplinary group of researchers outlined their vision for a new field aimed at integrating the work of engineers, who invent, train, and develop ML, with the work of social scientists and behavioral researchers from biological sciences. The umbrella term Machine Behavior was adopted by a large group of researchers advocating for an ethological approach to explaining AI (Rahwan et al., 2019). In biological ethology, researchers aim to understand animal behavior at four non-exclusive levels: function, mechanism, development, and evolution (Tinbergen, 1963). For example, one might explain mammalian herding as a learned behavior (bison herd because they find social contact rewarding) or as the result of natural selection by predation against loners (bison herd because solo bison fail to reproduce). In Machine Behavior, Rahwan and colleagues (2019) argue XAI would benefit from an adapted ethology for machines. For example, we can explain behavior as a function of the model's development (e.g., behaviors caused by idiosyncrasies in the training data, or the way feedback is incorporated into behavior), function, evolution (e.g., market forces or hardware capabilities that selected for certain types of algorithms), or mechanism (e.g., knowledge about architecture from under the hood). Our proposal for an experimental approach to XAI inspired by the study of human cognition can be nested within the ethology of Machine Behavior, and can similarly be leveled at different types of explanations. We can conduct experiments that explain behavioral consequences of being trained for certain tasks (development; e.g., fovea-like processing filters emerge naturally from learning to attend in visual scenes, Cheung et al., 2016) or experiments that identify behavioral consequences of different architectures (mechanism; e.g., visual crowding occurs in basic deep convolutional neural networks, but not eccentricity-dependent models, Volokitin et al., 2017). The study of human cognition has similarly found ways to explain behavior at different ethological levels (e.g., structure vs. function, Nairne, 2011). In summary, the experimental approach we are proposing is a natural extension of Machine Behavior and other recent efforts in XAI to benefit from behavioral science, and we expect this field to gain popularity at a rapid pace. It differs from other efforts at XAI in the emphasis on inferring cognitive models and causal explanations from experimental methods, analogous to the way cognitive psychologists study human behavior.

There is already a small group of researchers in machine learning who are doing the type of work for which we advocate. These are machine learning and computer science researchers who have taken inspiration from insights and methods popular in psychology and are actively publishing in major ML and computer vision conference venues (e.g., NeurIPS, ICML, ICLR, CVPR), so researchers who are considering whether a science of machine behavior is a viable path to publishing should not be dissuaded. We review this work extensively in "Artificial cognition", and provide a framework for using their efforts as a model going forward. Given that psychology is the science of behavior, broadly construed (including non-human animal behavior, computational modeling, etc.), there is no reason that work in Artificial Cognition should not also be published in psychology journals.

A review of XAI for psychologists

In this section, we provide a selective review of XAI that is intended to be accessible to readers without a background in ML and to illustrate the strength of a psychology-inspired approach to explanation. This section is meant to establish the case that the current state of XAI has a blind spot that artificial cognition could help fill. Readers already convinced in the case for artificial cognition can advance to "Artificial cognition". Note that there is a large family of explainable AI models that do not incorporate a 'deep' architecture. Models like linear regression, decision trees, and many types of clustering describe the relationships between variables in mathematically simple terms: either as a linear or monotonic function, or as the consequence of some quantitative decision threshold or rule-based logic (Molnar, 2019). While these models are interpretable, in many realistic settings their performance does not compete with deep neural networks (DNNs). There is a tradeoff between performance and model complexity (Gunning, 2017), matching the intuition that deep learning models perform well because they capture complicated relationships that cannot be easily represented. We explicitly limit our review to XAI for DNNs, which are the focus of most XAI research given their opaque nature, exemplary performance, and market excitement.

XAI research has undergone a recent explosion, slightly delayed but concomitant with the rise in popularity of deep learning's black boxes (see Fig. 1).

We cannot review the entirety of XAI methods, but we have selected a representative sample of state-of-the-art techniques for explaining predictions made by DNNs. For a more exhaustive review of XAI, we refer interested readers to several excellent, more thorough reviews, some of which we have drawn from for this section (Gilpin et al., 2018; Gunning & Aha, 2019; Guidotti et al., 2018; Mueller et al., 2019; Molnar, 2019; Lipton, 2017). Notably, each of these reviews presents a slightly different taxonomy for categorizing approaches to explanation (Hoffman et al., 2018b). For example, some emphasize the importance of explanations for global versus local behaviors, whereas others categorize XAI techniques based on whether they are model-agnostic or model-specific. These are all reasonable approaches, but note that there is a good deal of overlap between categories (Guidotti et al., 2018). Our goal is to advocate for an experimental psychology-inspired approach to generating explanations, so our taxonomy tactically categorizes approaches to XAI by what they *can't* explain or do, emphasizing the strength of our proposed approach. Note that we believe existing XAI approaches can be powerful, intuitive, and ingenious, and we are not recommending that they be abandoned in favor of psychological methods. We are merely arguing that current best practices possess a blind spot that is complemented by an experimental approach inspired by cognitive psychology. Specifically, XAI currently lacks the practice of causal inference by the attempted falsification of *a priori* hypotheses.

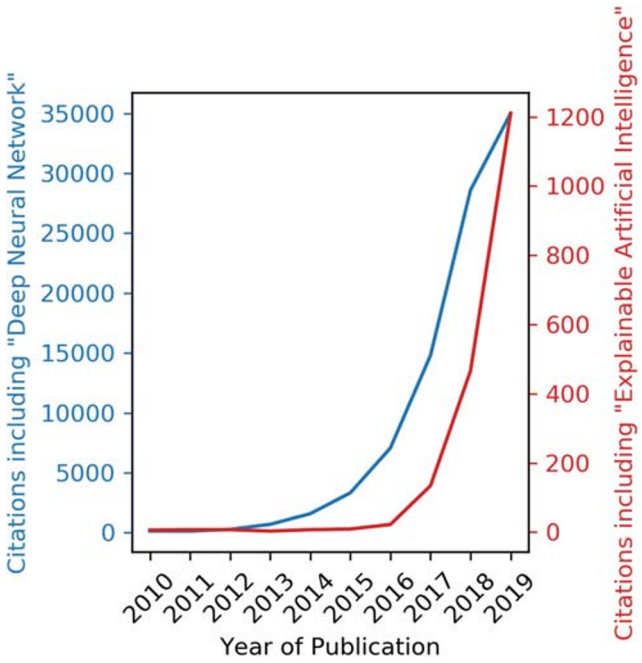

Fig. 1 Citations by year, via Google Scholar search for the quoted search terms. Deep learning and XAI have grown at a similar rate, although the deep learning research corpus is an order of magnitude larger and had a head start. Note the difference in the scale of the red and blue axes. Citation query occurred on January 10, 2020

Proxy models

An intuitive approach to explaining DNNs is to create an adjacent model with an interpretable architecture and to build it such that it makes similar decisions as the black box of interest. We then use the simpler model's interpretable decisions as a proxy for the black box's decision-making process. The simplest case of this would be using a linear regression, generalized linear models, or generalized additive models, which are very interpretable, to approximate a more complicated function. We do this all the time in psychology when we model behavior, which is no doubt a non-linear, non-convex function. As a general rule, these models sacrifice some predictive power in the service of explainability (Gunning, 2017).

If the approximation made by the linear proxy models is too severe, we can train the proxy network to convert the NNs choices into a decision tree. This effectively distills a neural network's approximated function into a series of conditional if-then rules. With a notional idea of the neuronal receptive fields in a model, we can distill the activity of a neural network into an and-or decision tree that explains how a class was chosen; the and-or graph learns the activity-related contingencies from a pre-trained DNN (Si & Zhu, 2013). At each level of such a hybrid DNN/decision tree model, we can observe what alternatives each node of the model is choosing between, and infer its purpose. For example, within a network that classifies handwritten digits, a node with possible outputs "3" or "8" responds to information that would join the ends of the 3 (Frosst & Hinton, 2017). Even without the ability to inspect a neuron's receptive field, abstract rules can be extracted automatically and pruned into a more digestible representation of the larger network (Zilke et al., 2016). By ditching unnecessary nodes and connections, pruning can result in a simplified network, more conducive to explanation. The critical idea is that neural networks contain highly efficient subnetworks that perform similarly to larger, more redundant networks by virtue of a lucky random initialization. Discovering them within a network highlights simpler processes (Frankle & Carbin, 2019).

Introspective models

XAI has reached a point where researchers are training DNNs to interpret the decisions of other DNNs (Zhang et al., 2018). We refer to these types of explanations as

introspective[2] because they suffer from the same drawback as introspection for human behavior: The process by which we introspect might not actually capture the function of the mind—it also requires an explanation. These AI-for-XAI solutions provide appealing explanations but have the obvious drawback of adding an additional black box to the system, inviting more explanation. Another complication with such methods is that the workhorse of deep learning, supervised learning, is generally not an option. As we have remarked earlier, people have trouble explaining their own decisions, which makes obtaining the required "ground truth" explanations for training very difficult.

Notwithstanding, these techniques form appealing explanations. We might have the introspective network learn to associate semantic labels to the nodes of hidden layers in a visual classifier, or train an adjacent NN suited for natural language processing to attach verbal descriptions to a DNN for image classification. This might result in explanations like "this bird is a Kentucky Warbler because it is yellow..." Hendricks et al. (2018). Users can easily see what the introspective network has learned about the first black box. The process can be applied to reinforcement learning agents as well; the introspective network learns to caption the internal state representation of the agent using sentiment analysis (Krening et al., 2016). As noted by Lipton (2017), this introspective network is optimized for generating explanations according to previously observed ground truth from human agents, which may not faithfully reproduce the artificial agent's internal state.

Correlative techniques & saliency maps

Unlike proxy models, which sacrifice some predictive power to incorporate an interpretable architecture, or introspective models, which train a second black box to explain the first one, many correlative techniques paint a picture of black box processing as is, without altering model design. These methods display the relationships models have learned and in some cases can highlight the subset of information in the input the model's correlations are weighting heavily in its predictions. Users must be careful to avoid causal logic when using correlative XAI techniques. Specifically, correlative techniques are susceptible to the threat of multicollinearity: there are often many possible causes for a given correlation.

Partial dependence plots (PDPs) illustrate the importance of a feature or variable to a model by displaying its output in response to iterative perturbations to the given input feature. For example, we might take a parameterized model trained to predict likelihood of car accidents based on driver data, then ask it to make predictions for a dataset where the variable representing years of experience is set to 0 for all drivers, then again at 1, 2, and so on. The resultant predictions are collapsed into a function that describes the model's treatment of driving experience over the entire range of inputs. The technique is extremely flexible and can be applied to any ML model using tabular data, such as predicting risk of developing diabetes based on blood glucose levels (Krause et al., 2016). We categorize PDPs as a correlative method because the conclusions about feature importance are expressing something about the correlations the machine has already learned. This means any conclusions drawn from these explanations must be tempered with the same caveats we apply to correlative logic in psychological science: Partial dependence does not indicate that the perturbed variable causes the outcome in the real world, and multicollinearity in the training data, which seems likely in large data sets, can complicate interpretation. Moreover, as noted by Molnar (2019), because PDPs present the average predicted output across the range of perturbed values, they can hide heterogeneous effects.

One way to circumvent the challenge attached to potentially heterogeneous effects is to display not the mean of perturbed predictions but the entire distribution. This technique, called individual conditional expectation, plots partial dependencies for all instances of the perturbed input rather than the model's global prediction (Goldstein et al., 2015). A similar approach, permutation feature importance, illustrates the increase in prediction error of a model following the iterative permutation of its features (Fisher et al., 2018). A similar concept is employed in iterative orthogonal feature projections, where an n-dimensional dataset is reproduced n times by transforming n-1 input features into an orthogonal projection, essentially removing the nth feature's effect on the prediction. The resultant relative dependencies are ranked to list the contribution of different input features (Adebayo & Kagal, 2016). These methods all apply a causal logic by manipulating input features one at a time, but the ranked feature importance is still based on the correlations of an internally consistent model. In other words, these methods do not retrain on new data with input features selectively removed. They estimate which features the model cares about by displaying hypothetical predictions based on previously learned correlations; they allow us to see when a model may be unfairly over-indexing on protected attributes, like gender or race. The XAI techniques listed above are black-box techniques, meaning they can be performed without access to the model internals.

Perhaps the most popular correlative technique is the family of gradient and activation visualizations generally

[2]NB Computer scientists will sometimes use the term "introspective analysis" to refer to any process that examines internal activations, which is separate from what we are describing here.

referred to as saliency maps (Simonyan et al., 2013; Ancona et al., 2018). DNNs rely on the backpropagation algorithm to compute the gradients required for parameter updates during training. The error function that specifies the difference between a model's prediction and the ground truth of the labeled data also determines, through differentiation, the direction that weights need to change to improve performance. So at every layer of the model, including the highly interpretable input layer, we can visualize the relationship between those neurons' activity and the ultimate decision. This is known as sensitivity: the degree to which changing a neuron's activity would affect the model's outcome (Simonyan et al., 2013; Selvaraju et al., 2017). Saliency maps generate highly intuitive explanations, as one can "see" which elements of a stimulus are related to the decision. Our challenge to the adequacy of saliency explanations is that they point to a manifold of explanations, with no ability to specify the cause of a model's ultimate decision; many different combinations of data could produce the depicted correlation. Correlative explanations can be misleading when features exhibit high multicollinearity, which seems likely with the larger datasets that enable deep learning. Others have criticized saliency methods as being insensitive to model architecture and data randomization (Adebayo et al., 2018). The output of some popular saliency maps resembles the output of simple edge detectors, which require no training and no special data processing. The argument, then, is that the correlations highlighted by visualization techniques are neither diagnostic nor causal.

Post hoc explanations

Post hoc methods invite a human user to rationalize the cause of a model's behavior after having observed the explanation. Most XAI methods fall into this category (Sheh & Monteath, 2018), which is similar to how humans anecdotally rationalize their own decisions (Lipton, 2017), and are subject to the same flaws. Like correlative techniques, post hoc explanations do not sacrifice predictive power in the service of explanation. Unlike correlative techniques, they introduce some causal logic into the explanation in the form of systematic perturbations, deliberate bottlenecks, or other manipulations. Anecdotally, these explanations feel very compelling. The downside is that these methods invite the user to project their biases onto the explanation, especially the confirmation bias.

Post hoc explanations perform systematic perturbations to a stimulus and observe changes to the model's prediction. These perturbations can take the form of progressive occlusion (systematically applying a small mask to every part of the input; Zeiler & Fergus 2013), ablation (zeroing out the activation vector of neurons or layers in the model), sensitivity (iteratively removing segments of an image that produce the smallest decrement in the correct classification vector until the classification changes; Zhou et al., 2015), or noise-injected perturbations (Fong & Vedaldi, 2017). In the case of an image classifier, the resulting explanation is another image representing the subset of essential information required for the original classification. A very popular version of this approach is Local-Interpretable Model-Agnostic Explanation (LIME), which uses perturbation logic to generate linear, local explanations for the decisions of any classifier (Ribeiro et al., 2016). The technique approximates a learned model's complex decision boundary by zooming in to the point where a linear approximation can be made. Decisions in the neighborhood of the zoomed-in portion inform the generation of a linear rule via iterative perturbation. The technique can only inform local explanations, which (for image classification) take the form of images where any contiguous area of pixels not weighted heavily in the generation of the linear approximation is omitted. We have categorized LIME and other perturbation methods as post hoc explanations because they invite the user to rationalize why the remaining pixels cause the classification. For example, in a case explaining which pixels of a guitar-playing dog image contribute to the decision *acoustic guitar*, parts of the dog's face remain highlighted by the explanation. Even though the explanation highlights pixels in both the dog face and the guitar, we venture a guess that most human interpreters would not admit the dog-face pixels as part of their explanation. Additionally, there is considerable risk of injecting bias in defining the neighborhood of the local explanation: "for each application you have to try different kernel settings and see for yourself if the explanation makes sense" (Molnar, 2019). A related technique by the creators of LIME finds "anchors" for a classifier's decision, which are simple, interpretable rules that specify a local decision boundary by finding the subset of the feature space for which changes to other features do not alter the decision (Ribeiro et al., 2018). In Fig. 2, we used LIME to generate explanations for image classification that illustrate vulnerabilities to human interpretation that we expect are common to post hoc explanation techniques.

Attention in DNNs has gained popularity within ML research, allowing for a different approach to explainability. Like its human inspiration, attention allows neural networks to focus on a subset of the available information (Mnih et al., 2014) and has lead to widespread performance improvements across a variety of tasks. In short, attention models learn to generate a guidance map not unlike modern theories of human attention (e.g., Wolfe & Gray, 2007), that is based on data-to-data correlations in the input. Attention can be used as an explanatory mechanism because it tells us which information the model or agent weighted

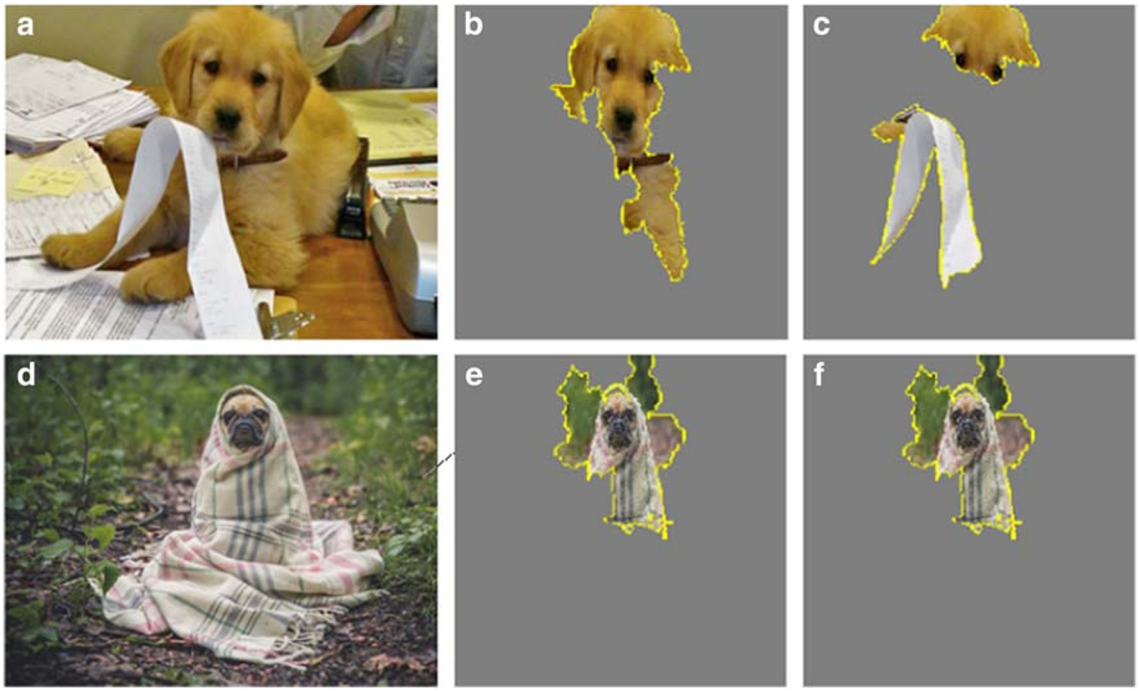

Fig. 2 **a** Input of a puppy doing taxes. **b** LIME's explanation for Inceptionv3's classification of "golden retriever". These are the superpixel areas that LIME has decided are most important for that class. The explanation matches our intuition because it contains parts of the input that include a golden retriever. **c** LIME's explanation for the class "toilet tissue". The long receipt is visually similar to toilet tissue, and we expect that most interpreters would be happy with this explanation. However, the explanation also includes a large area of the puppy's head. LIME determined this area is vitally explanatory, but we expect many observers would ignore this part of the explanation because it does not confirm the expectation of what an explanation for "toilet tissue" ought to include. **d** Input of a pug wrapped in a blanket. **e** LIME's explanation for InceptionV3's classification of "bath towel", its best guess for this difficult input. **f** LIME's explanation for InceptionV3's classification of "three-toed sloth", its fourth-best guess, which is identical to the explanation in (**e**). Does your confidence in the explanation for (**e**) change after seeing the explanation for (**f**)? These criticisms are not of LIME but of how we expect humans will interact with post hoc explanations

heavily (and which information it omitted) to perform its task (Mott et al., 2019). This process limits the range of possible explanations as it reduces the effective input space. In practice, this technique can be used to highlight the input an autonomous car relied on to update its steering angle (Kim & Canny, 2017), or to illustrate long-range dependencies that guide effective language translation (Lee et al., 2017). The bottleneck logic for attention as an explanatory tool has gained popularity, but critics have noted that the approach has some flaws despite its intuitive logic: learned attention weights—the values that 'guide' the attentional bottleneck—are often uncorrelated with gradient-based feature importance, meaning those areas of input do not drive the model's decision; and different attention patterns can lead to the same decision, questioning their diagnosticity (Jain & Wallace, 2019; Serrano & Smith, 2019). Even if attention gave us a completely reliable bottleneck—if nothing outside the window could influence model behavior—we would still be faced with the challenge of generating an explanation from within that narrowed range.

Another compelling post hoc rationalization is a class of a explanations that reverse-engineer exemplars of a neuron or layer's receptive field. This can be done using activation maximization techniques that find the maximum value of the dot product between an activation vector and some iteratively sampled image set (Erhan et al., 2009; Yosinski et al., 2015) or an iteratively generated image (Nguyen et al., 2016; Despraz et al., 2017). The maximal argument is taken as that neuron's receptive field, or preferred stimulus. The technique produces unmistakably identifiable yet surreal visualizations (Olah et al., 2018). These receptive-field exemplars can be further enhanced with human-annotated or automatically extracted conceptual information (Kim et al., 2018; Ghorbani et al., 2019). One shortcoming of these methods is that they only reveal compelling visualizations for concepts that are represented by single neurons, when some representations are achieved by distributed superposition over multiple neurons. Moreover, adversarial cases show us how images that elicit activation can mismatch human intuition about what such an image contains (Nguyen et al., 2015). There is growing evidence

that humans can identify adversarial images above chance in *n*-alternative forced-choice tasks, suggesting that humans can form intuition for how a computer vision model treats an adversarial example (e.g., Zhou & Firestone, 2019; Elsayed et al., 2018). Notwithstanding, there can be surprising heterogeneity among images that activate the same neuron as indicated by feature visualization, which calls their usefulness as explanations into question.

Example-based explanations

The above methods create explanations by trying to express how models and agents work. Example-based explanations find the limits of function and explain black box processing by illustrating ML constructs such as decision boundaries with counterfactual, adversarial, and critical examples as well as prototypical ones. The downside to these example-based explanations is that they generate these exemplars automatically from a massive stimulus space, so human-level interpretations of these explanations are *ad hoc*.

A simple way to gain intuition for a model's decision boundary is to see which inputs it treats as belonging to the same or similar classes. This is not causal logic, but it provides a gist of the shape of the categories the model has learned. A simple approach is to find the *k*-nearest neighbors or to follow another (dis)similarity metric (most commonly Euclidean). The intuition gained from exploring similarity space is vague, but powerfully intuitive (e.g., I can see these are all dogs, but what is it about dogs that the model is using to classify them as such?). But it also illustrates important failures when nonconforming inputs are found to be similar by the model. Nearest-neighbor algorithms follow an intuitive logic that can be baked into a more sophisticated example-generating approach; because it provides a measure of input homogeneity, it can be used by a deep network to learn to recognize nonconforming instances, including out-of-distribution test cases (Papernot & McDaniel, 2018). Nearest-neighbor algorithms feel deeply intuitive but fall short when expressing similarity in more than three dimensions. Instead, we can rely on the popular *t*-distributed stochastic neighbor embedding (t-SNE) method for capturing *n*-dimensional representational similarity in low-dimensional space (Maaten & Hinton, 2008). t-SNE solves the multidimensionality problem by mapping data points into a space conductive to visualization while preserving local similarity structure.

Deep learning is a powerful method for non-linear function approximation. The above methods are useful to express a convoluted decision boundary in low-dimensional space, but simplification can obscure some of a model's intricacies. One way to do this without dimensionality reduction is to present a user with inputs corresponding to local peaks in the output distribution. These prototypes present actual members of the dataset that are discovered using a clustering algorithm and are selected to represent the model predictions closely (Bien & Tibshirani, 2011). For example, with the MNIST dataset for hand-written numerals, we can use prototype methods to see which examples from the input space best represent the model's interpretations of the numerals 1-9. Alternatively, we can use a similar process to identify criticisms, which are also elements of the input space but are presented to the user when they do not fit the distribution of output labels to which they belong (Kim et al., 2016). In other words, criticisms present the user with outliers—cases where the model's behavior defies expectation.

Prototypes and criticisms are drawn from the input space, but we can also generate stimuli that tell us something about the model's decision boundary. Counterfactuals answer the question: How far do we need to perturb a stimulus before the model's decision changes? This can be achieved by iteratively perturbing an input feature until the decision changes, and minimizing the required change (Van Looveren & Klaise, 2019). The resultant stimuli tend to appear very similar to the original stimulus, and they impart a useful intuition for the decision boundary. Adversarial examples are similar, except the intent is to generate input that tricks a trained network into making mistakes (Szegedy et al., 2014). The adversarial input therefore tells us something about the decision boundary of the model under attack.

We are very excited by example-based XAI techniques, as they are epistemically aligned with the approach we advocate for in "Artificial cognition". We agree with the growing consensus that abductive and contrastive logic have a role to play in XAI (Hoffman et al., 2018a; Ilyas et al., 2019). By definition, these methods work iteratively and automatically, producing many examples of failure to form an idea of the boundary; they describe model function by illustrating its limits. So intuiting model behavior as a function of similarity, counterfactuals, adversaries, and criticisms involves interpretation with many observer degrees of freedom. In contrast, we advocate for an *a priori* approach, where scientists falsify select hypotheses rather than automatically generate many hypotheses.

Artificial cognition

Given the need for satisfying explanations to black-box behavior, recognizing the appetite for a science of machine behavior within computer science, and recognizing cognitive psychology's rich history of developing models of the mind through experimentation, we advance a hybrid discipline called *Artificial Cognition*, first coined by Ritter et al. (2017). Artificial Cognition can be thought of as a

branch of the Machine Behavior movement toward XAI, unique in its emphasis on cognitive models that are inferred from data elicited via experimentation rather than directly observed. Psychologists may recognize the distinction as being similar to the historic transformation that psychology endured in the 1950s when early cognitivists objected to the contemporary dominant view that mental processes were inadmissible as topics of scientific study; psychology at the time was limited to behaviorism and its models of stimulus and response. New perspectives (e.g., Chomsky et al., 1959) gave cognitive psychologists permission to invoke mental constructs, opening the door to novel predictions and powerful theories of the mind. We can say things like "I was distracted because my attention was divided," or "I avoided the collision because the looming motorcycle captured my attention." These explanations are intuitive and imply causation, and lead to falsifiable hypotheses about human behavior. We can apply the same discipline to machine behavior.

Benchmarking versus experimentation

As we have discussed, an important tool in ML research is benchmarking, where researchers measure the performance of different models or agents on standardized tasks. Benchmarking compares new models to the established state-of-the-art, and is an indispensable practice for engineers vying to produce the best product. However, benchmarking has also been wrongfully employed as a tool for explainability. We have observed cases with an epistemic logic of this format: If I can build it to be better, I understand it[3]. The reason benchmarking yields unsatisfying explanations is because a model or agent may be effective for reasons having nothing to do with the inventor's latest innovation. The logic is confirmatory, and qualifies as verificationism.

The reason experimentation produces satisfying explanations is because it is inherently falsifying. Cognitive psychology has a strong Popperian tradition of falsificationism (Popper, 2014), where we gain confidence not by finding confirmation of our theory but by defeating alternatives *against* it. A good explanation should have survived fair attempts to falsify it. It is at this point we think it may be helpful to describe what we mean by experimentation. We have observed that ML researchers sometimes use the word "experiment" to mean "let's expose an agent/model to an environment/input and see how it performs," or "let's see if our new agent/model can out-perform the state-of-the-art." When cognitive psychologists talk about experimentation, they are referring to attempts to falsify null or competing hypotheses to shape their theories. Our purpose is not to benchmark the state-of-the-art but to look for reasonable counter-explanations to our working theory, any one of which should be fatal if supported. A similar abductive approach has been called for by scholars advocating for cross-disciplinary approaches to XAI (Klein, 2018).

A framework for explanation in artificial cognition

What should research in Artificial Cognition look like? Artificial Cognition can be a sub-discipline dedicated to inferring causal behavioral models for AI following a domain-general scientific framework. We suggest a research pipeline that identifies a behavior and its environmental correlates, infers its cause, and identifies its boundary conditions. These steps mimic the arc of research programs in prominent psychology laboratories and science in general. These should be followed by informed attempts to alter behavior by changing the machine according to our theory of the machine's mind, which is something that psychology researchers typically cannot do[4]. Once we have an idea of the cause of a behavior, we should be able to change that behavior by changing the machine. Cogent experimentation following this arc should produce satisfying, causal explanations of machine behavior. Moreover, this research arc avoids the common causal reasoning pitfalls reviewed in "A review of XAI for psychologists".

We would not presume to teach science to readers. The purpose of this section is not to re-invent the wheel but rather to highlight exemplary cases of behavioral science at work in XAI, often with methods adapted explicitly from cognitive psychology, for each step of the recommended scientific framework. This section, therefore, doubles as a review of exemplary cases of XAI using artificial cognition. Please note there is nothing unique in this framework to psychology—it is a domain-general scientific framework that should feel familiar to researchers from many disciplines. The reason psychology may be specifically adaptable to XAI is because of its tradition developing experiments for black-box models. Small groups of ML researchers have already been adapting psychology experiments to suit their black-box models, and we review these cases here as examples of how psychologists can continue this work.

[3] n.b. while we reject this logic, we note the negative form of this argument may be true: Feynman's famous "what I cannot create, I do not understand."

[4] This last step will likely require collaboration with ML researchers. Alternatively, it is becoming easier to learn and do-it-yourself with the proliferation of free online courses, improved and well-documented software frameworks, and the open-sourcing of most models/agents.

Document variations in behavior

First, we must identify behaviors and establish them as subjects of study by correlating them with changes in the machine's task, training, architecture, or environment. The effect must be real and it must not be random.

Standardized tasks are very helpful for these purposes. In cognitive psychology, dominant paradigms produce thousands of published studies every year, which are subtle variations on a relatively small set of tasks. ML researchers have also produced batteries of tests that are useful for assessing performance on a variety of tasks. The popular Arcade Learning Environment (Bellemare et al., 2013) offers RL researchers multiple Atari games with a desirably small action space (nine joystick directions with or without button press) and a variety of behaviors to learn. In supervised learning, a popular task for image classification is top-5 error rate on ImageNet (which forms the basis of the famous ImageNet Large Scale Image Recognition Challenge that, until 2017, benchmarked the state-of-the-art on an annual basis and produced a watershed moment for deep learning in 2012). Presently, ML researchers are developing batteries of tests that measure more general behaviors, which will be very useful for developing models of mind. For example, Deep Sea is a common test of an RL agent's ability to explore an environment efficiently. The travelling salesman task is one of the oldest problems studied in combinatorial optimization and many heuristic solutions have been developed. Yet it has been recently revisited in the context of learning agents (Vinyals et al., 2017). Furthermore, the new PsychLab package directly imports classic visual search tasks and other paradigms (e.g., Treisman & Gelade, 1980) into an RL environment to measure elements of visual cognition (Leibo et al., 2018). Access to a variety of well-understood paradigms is an important step toward identifying curious behaviors that will form the subjects of inquiry for artificial cognition.

A good example of using standardized tasks to identify a behavior worth explaining is Behaviour Suite for Reinforcement Learning (bsuite, Osband et al., 2019). The authors documented the relative performance of three different RL agent architectures (actor-critic RNN, bootstrapped deep Q-network, and deep Q-network) on tasks that intuitively measure different faculties (e.g., the *Memory Length* experiment to measure long-term memory, *Deep Sea* for exploration). The researchers correctly predicted that a bootstrapped deep Q-network would exhibit specifically superior exploration in a new environment, whereas the A2C RNN performed better on the memory task.

This exploratory approach to documenting variations in behavior can also be used to identify vulnerabilities and areas to improve DNN functionality. Using three well-known DNNs (ResNet-152, VGG-19, and GoogLeNet), Geirhos et al. illustrated two important shortcomings in deep neural networks' ability to generalize (Geirhos et al., 2018). Whereas humans can maintain perception through different visual noise distortions, they showed that DNN accuracy for most of the distortion categories rapidly dropped off with increasing signal-to-noise—more so than the humans in their study. More interestingly, when the machines were trained on datasets that included one or more of the 12 distortion types, they only gained robustness against those particular distortions. This second investigation suggests that the DNN vulnerability to noise distortions is not simply a consequence of the high-quality images in datasets we normally train DNNs with. If that were the case, then re-training the DNNs with the various noise assaults should have conferred some protection against other distortions. Instead, across the board, the DNNs learned robustness only against the specific distortion present in their training set.

These examples illustrate how researchers should begin the path to explainability by documenting changes in behavior that correspond to differences in task or stimulus. Other promising starting points involve correlating behavior with changes in data composition (e.g., identifying bias emerging from supervised learning between different data sets), learning algorithm, architecture (e.g., establishing that A2C and b-DQN RL agents excel at different tasks Osband et al. 2019), or, in RL, changes in environment (e.g., showing that introducing additional obstacles and environmental features in a game of hide and seek causes agents to learn surprising fort-building behaviors; Baker et al. 2019).

Infer the cause

Correlations are insufficient to infer the existence of black box processes that cause behavior. Instead, hypothesized processes must survive a fair test of falsification by exposing the machine to carefully controlled circumstances designed to rule out alternative explanations. The experimenter must curate sets of stimuli, environments, or agents that have and do not have the variables hypothesized to be affected by the proposed process; all other variables between stimulus sets must be controlled. Designing experiments for behavior should be familiar to cognitive psychologists, but there are some critical differences between experimenting on machines versus humans, which we discuss in "Getting started".

We have seen some excellent cases of falsifying experiments in ML research. *PilotNet* is NVIDIA's neural network-based steering algorithm for autonomous vehicles,

which takes images of the oncoming road as input and outputs steering angles (Bojarski et al., 2017). Its creators wanted to understand what features in the visual input caused PilotNet to vary its output. The activity of neurons in the higher layers of their network correlated strongly with features in the input images that they anecdotally identified as ground-level contours. Recognizing that the correlation does not necessarily imply a causal relationship between those features and steering angle output, the researchers formalized their observation into a testable and falsifiable hypothesis by subjecting PilotNet to an experiment. Input images were modified three ways: (a) displacing pixels that were identified as salient by their correlative technique; (b) displacing all other pixels, or (c) displacing the entire image. If the authors' correlative technique was incorrectly identifying crucial visual features for steering, then we would expect similar performance between comparisons of [a,b], [b,c], and/or [a,b,c]. The data showed that a and c were similarly responsive in steering updates, with condition b making minimal adjustments, failing to reject the alternative hypotheses and lending confidence to their explanation that ground-level contours identified by their saliency algorithm cause changes in PilotNet's steering.

We have also seen cases where causal explanations are inferred from a model's behavior under competition between controlled features within a stimulus (rather than between stimuli in different conditions). Seeking an explanation for why CNNs for object recognition can identify inputs without difficulty using only local texture clues—where the global shape is completely destroyed (Brendel & Bethge, 2019), Geirhos and colleagues conducted a clever experiment that directly pitted global shape and local texture cues against each other (Geirhos et al., 2018). The logic of their manipulation mirrored many experiments in visual cognition, where we present the observer with a stimulus containing two conflicting features and see which one the visual system 'prefers'—such as in Navon's ambiguous stimuli containing both global and local structure (e.g., a big arrow pointing left made up of little arrows pointing right) (Navon, 2003). In this case, Geirhos created stimuli using a style-transfer algorithm that applied one input's texture to another input's shape, resulting in chimeric stimuli, such as a cat-shaped image with elephant skin texture. The question then was whether the DNNs classified the inputs as belonging to the shape-defined class (cat) or the texture-defined class (elephant). Humans in their study exhibited a strong shape bias, consistent with the psychology literature (Landau et al., 1988), but DNNs tended to classify the chimeric image as belonging to the texture's class. Models retrained on the chimeric style-transferred images were more robust against the texture bias, as they were more likely to categorize using shape than a standard ImageNet-trained model. Because the behavior changes following retraining on carefully crafted datasets, we can attribute the predominant texture bias in CNNs to regularities in ImageNet training data rather than some principle innate to their design.

Another exemplary case of experimentation in ML research comes from Ritter et al.'s investigation into shape bias in one-shot object recognition (Ritter et al., 2017). Those authors, who coined the phrase *Artificial Cognition* and advanced the use of insights from cognitive psychology for understanding ML, devised an experiment to determine whether DNNs would exhibit the same biases as humans when learning new objects. Specifically, they hypothesized that shape, rather than color, would be a biasing feature when generalizing from known to novel objects. They borrowed stimulus sets modeled after tasks from developmental psychology laboratories that feature novel objects that can vary in shape or color but control other major variables such as background (Landau et al., 1988, 1992). They also fashioned their own stimulus set following these principles to generalize to more naturalistic stimuli (as opposed to novel nonsense stimuli). The model in this case was pre-trained with a state-of-the-art image classifier to provide basic feature detection. The model was then asked to identify the most similar image from a novel support set that included shape-matching and color-matching stimuli. Results showed that the model was much more likely to identify the shape-matching novel object as belonging to the same category, confirming a human-like shape bias in object identity learning. The crucial design feature was the perfect experimental control provided by the shape- and color-matching probe, isolating the causal variable.

Investigations using psychology-inspired experimentation have also been used to discover properties of Gestalt perception in neural networks (Kim et al., 2019). These researchers asked whether DNNs for image classification exhibited the law of closure (Wertheimer, 1923). They hypothesized that closure would emerge spontaneously in neural networks trained on natural images. The authors examined how different networks trained under carefully controlled conditions would process pairs of triangle stimuli. Critically, the pairs were a complete triangle and either an incomplete triangle with Gestalt closure or an incomplete triangle without closure due to the rotation of its vertex elements. The networks' response similarity was greatest for complete and illusory triangles when the network had been trained on natural images versus networks with random weights or networks trained on random visual stimuli. This is an exemplary experiment with clear predictions, independent and dependent variables, and conclusions.

Many of the studies reviewed here reveal compelling explanations for AI behavior by tinkering with the model.

But how should we approach XAI in cases where, whether for practical or proprietary reasons, we cannot look inside the black box? We recently showed that it is possible to conduct XAI using techniques from cognitive psychology from completely outside the black box (Taylor et al., 2020). One common question faced by XAI researchers is how and where does a model represent different hierarchical features? A common approach to answering this question is to visualize the receptive fields of neurons throughout a model (e.g., Olah et al., 2018), but this is only possible with access to the model's internals. We reasoned that models using conditional computation—the ability to short-circuit the forward pass, making an early exit at an intermediate classifier—should show response time effects in their response corresponding to feature space. We theorized that models with early exit architecture should exhibit faster response times for stimuli that can be classified using features that are composed in earlier layers. In a proof of concept, we showed that MSDNet (Huang et al., 2017) was much faster to classify stimuli from ImageNet versus the much more challenging ObjectNet test set (Barbu et al., 2019), which was explicitly created to contain a less homogeneous and more intricate feature space. In a more controlled second experiment, we showed that MSDNet is sensitive to the statistical regularities in object-object and object-scene relationships that populate stereotypical scenes resulting in scene grammar effects in humans (Võ & Wolfe, 2013). The SCEGRAM test set (Öhlschläger & Võ, 2017) presents 62 scenes in four different versions, one with consistent scene grammar, and one with inconsistent semantics, syntax, or both. Building on the finding that NNs for scene and object recognition struggle with visual inconsistencies in SCEGRAM (Bayat et al., 2018), we found that, like humans, ANNs with early exit architecture exhibit RT effects in classifying objects across the four conditions. The ANOVA revealed that MSDNet was specifically challenged by semantic inconsistencies, which makes sense given that it is trained for object recognition. This means that on average the features required to correctly classify objects in scenes with inconsistent semantics could not have been accessed by the earlier classifiers. We were therefore able to make inferences about the relationship between feature space and model depth without inspecting model internals.

So far we have reviewed experimentation for computer vision, which represents our bias as vision researchers, but artificial cognition can be equally effective in explaining model behavior for other tasks. Gulordava et al. asked whether recurrent NNs for natural language processing learn shallow patterns between word frequencies or whether they learn grammar (Gulordava et al., 2018). They showed that their RNN made reliable predictions for difficult long-term dependencies in four languages. In a throwback to the cognitive revolution, they showed this was true even for sentences where meaning had been divorced from grammar, as in Chomsky's famous nonsense sentence demonstrations (e.g., "colorless green ideas sleep furiously; Chomsky, 1957). The experiment shows that RNNs with long-term memory can extract grammatical rules from simple text-based training data without any prior toward or explicit instruction to attend to syntax.

The process of designing and deploying experiments to infer the cause of machine behaviors is currently underutilized. We contend this represents a major opportunity for cognitive psychologists to contribute to a growing field with a hunger for experimentation.

Identify boundary conditions

If you believe you can explain when a behavior happens, then you should also be able to account for when the behavior stops. Identifying the boundary conditions of a behavior is an important element of explaining it, because it narrows the range of viable alternative explanations. More practically, it helps describe when a model or agent is likely to change its behavior, which is important for users, regulators, insurers, and developers. For example, an autonomous vehicle may perform above human level under normal conditions, but how does it perform under low light, in snow, or through road work?

Broadly speaking, there are two ways to establish the boundaries of an effect in psychology. One is by adjusting the intensity of a stimulus to titrate the level required to elicit an effect using psychophysical techniques (Macmillan & Creelman, 2004) or tuning curves. This approach delivers characteristic functions that describe how and whether a person perceives a stimulus given a range of intensities. The other approach is to expose the subject to a range of conditions with controlled alternatives, asking whether differences in behavior emerge under different circumstances. In the example above, we might compare an autonomous vehicle's performance on the same road under two different conditions, asking whether it behaves similarly, rain or shine.

The use of psychophysics to characterize an effect's boundaries is beginning to take hold in XAI (Rajalingham et al., 2018; Scheirer et al., 2014). An illustrative case that uses both boundary-defining techniques described above is RichardWebster et al.'s recent release of PsyPhy, a visual psychophysics package implementation available in Python (RichardWebster et al., 2019), which has already been applied to characterize the boundaries of different facial recognition software (RichardWebster et al., 2018). In that study, the authors psychometrically evaluated the

item-response curves of five different face recognition algorithms (and human performance) under different degradation assaults, including Gaussian blur, contrast, pink noise, brown noise, and variable facial expression (e.g., degree of emotion or degree of eye-blink), resulting in a detailed description of the models' abilities and faults. The detailed item-response curves illuminated some surprising conclusions, too. The authors originally predicted that deep learning CNNs would be uniformly superior to shallower networks or models with handcrafted feature detectors. The psychophysical approach showed that VGG-Face, a type of deep CNN designed to process images of faces, was the best algorithm in most circumstances for most intensities, but for a wide swath of the stimulus space, the handcrafted feature detector OpenBR outperformed several other deep CNNs. The psychophysical approach also revealed the surprising discrepancy between OpenFace and FaceNet, which share architectural similarities: both are variants of the same CNN architecture originally developed by Google. This empirical discovery prompted the researchers to propose that differences in the training sets used by these sibling models might explain the difference in performance. In short, the use of psychophysics prompted the discovery of several surprising boundaries and caveats to the face recognition algorithms' abilities.

Another approach to quantifying boundary conditions involves introducing systematic perturbations and measuring associated decrements in performance. This can be done with artificial perturbations such as deleting swatches of input or scale transformation (Zeiler & Fergus, 2013), or with naturalistic perturbations such as introducing visual clutter (Volokitin et al., 2017). In the latter case, researchers produced detailed functions describing the relationship between target eccentricity, flanker similarity, and flanker separation, much in the same way that humans experience visual crowding (Whitney & Levi, 2011). They concluded that targets and flankers in the input were often grouped by pooling functions often used throughout ANN processing hierarchies. Moreover, they were able to identify a combination of conditions required for robustness against crowding in a DNN trained on images without clutter: an eccentricity-dependent DNN where receptive field size of convolutional filters varies (classic vanilla DNNs do not have this multi-scale feature), with targets centrally-presented (like a human fovea), and spatial and scale pooling only at the end of the model's hierarchy (as opposed to throughout). This type of explanation, based on a psychophysical exploration for the boundaries of successful performance, is desirable in cases where we want to understand how and when a model will fail or succeed in naturalistic scenarios, which is a top priority for explainability research and responsible deployment of AI.

Toy with the brain

Finally, we recommend a fourth step in the research program arc that is normally not possible with humans. Machines offer us the freedom to alter neural networks. Showing that selective tinkering corresponds to falsifiable predictions for subsequent behavior provides a strong test of any working explanation, and should be regarded as the highest level of explainability. Psychologists do not have the ability to alter their black box for obvious ethical reasons[5], so this is an exciting opportunity for psychologists to test the strength of experimentally derived models of mind. For ML researchers and creators, this level of explainability describes next-steps for development. Having strong predictions about how the machine should behave under different alterations can lead directly to improvements in design and performance.

For a strong example of toying with the brain in ML, consider Leibo et al.'s deployment of their *UNREAL* RL agent in the newly designed PsychLab environment (Leibo et al., 2018). Those authors imported some classic visual cognition experiments into an RL environment with the hope of characterizing machine behavior against well-studied human phenomena. While UNREAL outperformed humans on many tasks (including exhibiting efficient visual search for feature conjunctions!), it had worse visual acuity than humans. Consequently, the authors hypothesized that UNREAL would preferentially learn large objects at the expense of smaller objects in the environment, and they confirmed this with an experiment that asked UNREAL to perform a localization task in the presence of small or large distractors. Results showed that UNREAL learned the task, which was to point to the object on the left side of space, slower in the large distractor condition. This finding confirmed their hypothesis that UNREAL's low acuity causes it to repeat the surprising and undesirable behavior of identifying the larger object, rather than the leftward object which was actually more rewarding. Critically, the researchers used this experimentally derived causal insight to further develop UNREAL. Rather than allowing UNREAL to view the entire field, which is computationally expensive at higher resolution input and is also the likely cause of UNREAL's fixation on large objects, they programmed an input processing filter inspired by the human fovea. Doing so increased performance on tasks in dynamic environments with higher input resolution without the prohibitive cost of distributed weight sharing. In short,

[5]Unless in animal or computational models, or in rare circumstances through drugs, transcranial magnetic or direct current stimulation, deep brain stimulation, or as a corollary to medically necessary surgery.

experimentation lead to a theory of mind for UNREAL's visual cognition that directly inspired better AI.

Another approach to manipulating a network in the service of explainability is retraining an identical model with different data. This affords testing hypotheses relating to the composition of data-dependent features. In order to explain the superiority of visual acuity for horizontal and vertical information in artificial and biological neural networks, Henderson & Serences tested whether ANNs would learn a similar bias and whether that bias depended on statistical regularities in the training datasets (Henderson & Serences, 2020). First, they showed that their pre-trained ANN did exhibit a bias toward over-representing cardinal orientations by measuring the distribution of tuning centers across neurons. The cardinal bias was correlated with layer depth. This was contrasted with the same distributions for a randomly instantiated version of the same model—a control condition designed to rule out the possibility that the orientation bias is incidental to the model's parameterization. Next, the authors re-trained VGG-16 on orientation-shifted versions of ImageNet, predicting that the cardinal orientation bias would match the perturbations on the training sets. In a compelling demonstration, they showed that the new models' orientation bias perfectly matched the shifted datasets' perturbation, which offers a tidy explanation for why neural networks develop these preferences.

Advantages of artificial cognition

Artificial cognition is model-agnostic

Experimentation, in the hypothesis-falsifying sense, is model-agnostic. We can seek explanations through experiments on any ML algorithm, whether it is supervised or unsupervised, or an RL agent, or a hybrid of the above. All the experimenter needs is the ability to manipulate input (stimuli) and observe output (behavior). Knowledge about the model's architecture is advantageous, in the way that knowledge from neuroscience can guide the space of feasible hypotheses in cognition, but not strictly necessary.

Artificial cognition does not constrain design

Explanations from Artificial Cognition do not come at the cost of constraining design. As we have seen with some XAI methods, more interpretable architectures sacrifice performance in the service of explainability (Gunning, 2017). All conclusions are drawn from outside the black box. Cognitive psychologists have learned to explain behavior without changing the software because we cannot.

Artificial cognition tests hypotheses a priori

Artificial Cognition is hypothesis-driven, resulting in *a priori* rather than *post hoc* explanations of behavior. Many of the explanatory mechanisms we have seen in the XAI literature are in the form of visualizations that correlate neuronal activity with a decision likelihood, or visualizations that display decision likelihood as a function of local perturbations (Ribeiro et al., 2016). These visualizations are a powerful tool in forming intuition for how certain nodes and layers interact with the data, but we must recognize that these intuitions form post hoc, posing a risk of confirmation bias (Wason, 1960). We urge caution when interpreting correlative visualizations, as they are not causal. We are also wary of automated perturbation-based explanations, as the causal mechanism is always inferred ad hoc. That is not to say that these tools are not useful in the service of explanation: They are indispensable in that they point toward early theories and hypotheses. In contrast, truly experimental methods employ falsifying logic and produce causal explanations, which align with how people explain events in the world (Sloman, 2005) and have been shown to produce more satisfying explanations of AI behavior (Madumal et al., 2019).

Caveats

Mechanisms of behavior & cognition

We might reasonably be concerned that these methods will not uncover the mechanisms underlying AI behavior, but instead describe behavior as a function of various inputs and patterns of learning, not unlike behaviorism. But, to the extent that cognitive psychologists have uncovered said mechanisms, artificial cognition should produce similar results without issue. We say this confidently because the process is identical: curate experiments, observe behavior, and infer the underlying processes.

Cognitive psychology deals in mechanisms all the time. Covert orienting is described as a mechanism of attention (Posner, 1980). The inhibitory processes in retrieval is described as a mechanism for remembering things stored in long-term memory (Bjork, 1989). Cognitive explanations give us useful metaphors to understand behavior. They are decidedly not behavioristic because the behaviorists famously forbade the consideration of mental constructs, such as long-term memory.

We can object to the critique on even stronger grounds. There are cases where psychology experiments can be used to uncover properties of the structure of a ML model. We recently applied response time analyses to a model designed

for object recognition. We wanted to make inferences about how the model processed semantic features using the SCEGRAM database (Öhlschläger & Võ, 2017). Examining the RT distribution in the model's response revealed five clear RT plateaus, which could only occur if the model had exactly five intermediate classifiers and an early exit architecture (Taylor et al., 2020). This illustrates how black-box experiments can reveal both cognitive-type analogies for the mechanisms of behavior and also the physical neuroscience-type mechanisms of the machine.

On transferring principles of human cognition to AI

It is vitally important that we are not interpreted as saying that the discoveries of research on human cognition should apply to AI. Unless the AI in question is explicitly modeled after the human brain and/or mind, there is no reason to expect that it should exhibit similar behaviors (and even then it might not). Artificial Cognition will work because the *methods* from psychology are good for developing black-box explanations for behavior. These are methods for learning about cognition, broadly construed—not just human cognition.

Ecological validity in human and artificial cognition

Psychology has been appropriately criticized for making generalizations about behavior from overly artificial settings and stimuli (Kingstone et al., 2008; Neisser, 1978) and unrepresentative populations (Henrich et al., 2010). These charges have been met with calls for an increase in applied research (Wolfe, 2016) where findings from psychological science inform human behavior in the real world. It would be counterproductive if artificial cognition inherited psychology's controversies along with the strengths of its method. As such, we encourage newcomers to take advantage of the "fresh start" in the study of behavior for machines and be mindful of the pros and cons of investigating behavior with different levels of ecological validity. Naturalistic stimuli will be vital to explain how models ought to behave in their intended use. Contrived stimuli will be important for anticipating edge cases and informing thorough models of behavior.

We speculate that it ought to be easier to achieve ecological validity with machines. It is a challenge in human psychology because research in laboratories is more tractable and grants us precious control. In AI, the naturalistic task the model is created for could more feasibly be perfectly recreated during experimentation (at least compared to human explainability research).

Getting started

In this section, we describe some practical considerations in making the transition from research on human cognition to research on artificial cognition, point to useful resources in getting started in XAI, and provide a concrete walkthrough from our own empirical work. We hope it can provide some useful insights for behavioral scientists studying machines for the first time.

Educational resources and useful code

Producing performant AI is the domain of ML research, where aspects of model design such as architecture, loss functions, learning and optimization algorithms are studied extensively. Reviewing these topics is outside the scope of this article, but a cursory understanding (at least) of ML principles is required to be making inferences about how machines make decisions, in the same way that neuroscience is part of every education in human psychology. To that end, we will recommend some resources here for psychologists to familiarize themselves with the principles of ML. Luckily, ML has a strong culture of making educational resources free and open-source.

Deep learning involves an intimidating mixture of mathematics drawn from linear algebra, calculus, and statistics. Fortunately, the entry-level requirements for understanding basic DNNs is a high school or undergraduate level understanding of each domain, common to many bachelors degrees in science. And instead of learning or refreshing each discipline independently, you can refer to helpful open-source texts[6], and intuition-building video lectures[7]. Before conducting research on DNNs, psychologists should understand linear transformations and basic matrix operations from linear algebra, the chain rule in multivariate calculus for understanding backpropagation[8], and gradient descent and regression from statistics, with which psychologists are already familiar. This is an egregious oversimplification of the math involved in deep learning, but those are the basics. Deep dives into academic material on deep learning are also available through high profile open courseware[9]. For

[6] e.g., Deisenroth et al. (2020), available at https://mml-book.com/

[7] e.g., 3Blue1Brown's inimitable YouTube series

[8] NB this is probably the most intimidating material but you also probably learned the chain rule in high school

[9] e.g., Stanford's CS231n at http://cs231n.stanford.edu/syllabus.html covers visual recognition architectures, which are the most heavily explored type of model in XAI

a math-based review of XAI methods specifically, consult Molnar's excellent ebook[10] (Molnar, 2019).

We don't expect that most psychologists will gravitate toward building new AI—that is the domain of ML researchers. Instead, artificial cognition will seek to explain extant models' decision-making, and to that end we will be experimenting with other creators' work. This means the bulk of the artificial cognitivist's work will require getting their hands on popular models and datasets. Fortunately, it is relatively easy to download pre-trained versions of most popular models from the literature for research purposes. Continuing ML's culture of open-access, when new breakthroughs are achieved they are often uploaded to public repositories for other researchers to replicate results. The most common repository is GitHub, and PapersWithCode provides a clean and searchable directory connecting publicly available models with the research papers that report them. Finding popular datasets for machine learning is also facilitated by open-access repositories, including Google's new dataset search[11]. Machine learning researchers are partial to the practice of using "awesome lists", which are lists of resources in a GitHub readme tagged with a subject and the word "awesome"[12]. We do not know the whole story behind the practice, but it certainly simplifies searching GitHub.

Finally, the vast majority of research in AI is conducted on models and agents written in Python. Many psychologists will already be using Python to program their experiments and analyses, but we suspect the majority are using MATLAB or a GUI-based environment to create their experiments and MATLAB, SPSS, or R for analyses. Python is completely essential to interact with deep learning models. Users migrating from MATLAB will appreciate the similarity of the Spyder[13] environment for writing Python, which is available through the free and open-source Python distribution platform, Anaconda[14]. Transitioning to Python is relatively easy for anyone who has already taught themselves MATLAB or R, and there are useful guides specifically for this conversion; NumPy, the main package for scientific computing with Python, hosts a useful guide for converting from MATLAB[15]. The Scipy Lecture Notes[16] are a free and evolving entry point to the Python for scientific computing ecosystem, organized into a series of self-contained tutorials.

We also want to share two hands-on tutorials that show how to interact with DNNs. The first is a simple python script that loads a pre-trained DNN for visual classification (InceptionV3), specifies a function for converting any image into a readable format by the model, performs a forward pass using any image as input, and returns the model's decision. Clone the linked repository[17], open and run the Python script in Spyder or the environment of your choice.

The second tutorial is more advanced, describing an experiment from start to finish, including loading large datasets, building, training, and testing a complete model[18]. This code describes the process of training, validating, and testing MSDNet for the response time methods study we described in "Artificial cognition" (Taylor et al., 2020). The output from this code is saved as a `.csv`, which is readable by any analysis software. Our analysis was written in Python using the NumPy and SciPy packages, and is practically indistinguishable from interacting with output from human subjects in various spreadsheet formats (e.g., `.mat` or `.sav`); no additional training or guidance is needed for psychologists. The output contained RT, classes, and confidence from each intermediate classifier. For 100 steps of confidence form 0 to 1, we found the minimum speed at which the model would confidently supply an answer. Responses were pooled across images from different experimental conditions (ObjectNet vs. ImageNet in Experiment 1; 2x2 factorial ANOVA for scene grammar conditions in Experiment 2) and analyzed using statistical packages in SciPy.

Considerations for analysis of artificial minds

One of the fundamental tools of experimental psychology is inferential statistics. Every psychology undergraduate receives extensive training in how to infer behavior in the population from patterns in a sample. The assumption underlying the pervasive use of inferential statistics is that there *is* a fundamental commonality between minds in the population: Our minds are alike, so your behavior in the laboratory informs my behavior in the world. Most psychologists would agree that the assumption is justified by the shared architecture of our brains, and is demonstrably robust against observation. However, the assumption of a common mind is impossible in artificial intelligence, where the diversity of "brains" is matched only by the ingenuity of computer scientists. It does not make sense to infer the behavior of one algorithm from the behavior of other algorithms with explicitly different architectures. Consequently, Artificial Cognition must dispense with the common use of inferential statistics, where measures

[10] https://christophm.github.io/interpretable-ml-book/

[11] https://datasetsearch.research.google.com/

[12] e.g., https://github.com/lopusz/awesome-interpretable-machine-learning

[13] https://www.spyder-ide.org/

[14] https://www.anaconda.com/

[15] https://numpy.org/doc/stable/user/numpy-for-MATLAB-users.html

[16] https://scipy-lectures.org/

[17] https://tinyurl.com/yxehuera

[18] https://tinyurl.com/y2pgwygk

of central tendency are compared between samples of individuals assigned to experimental and control groups, and where deviations from these measures are treated as statistical error to be minimized rather than signal to be analyzed. In contrast, Artificial Cognition will be more akin to the psychology of individual differences, where deviations from the mean represent true data. With humans, we often treat the error associated with different stimuli from the same group as a fixed effect in our analyses. For example with scene grammar studies in humans, all observations from a single individual are pooled across images from the same category; this data is expressed as a single mean entered into the ANOVA, with means for each level of the independent variables. In our RT study on scene grammar in CNNs, we instead studied the error attached to the 62 different scenes as a random effect. The resultant ANOVA allowed us to infer how MSDNet would treat unforeseen scenes drawn from a distribution with similar scene grammar. Additionally, we often take a measure of central tendency over many observations of identical stimuli to reduce the error in our observation. In the majority of cases, there is no need to repeat exposure to a pre-trained algorithm. Unless there is stochasticity embedded in the model, such as in RL agents programmed to explore their environments, it will react to identical stimuli the same way each time.

In addition to the re-purposing of inferential statistics, we can quantify AI behavior with techniques developed to characterize individual differences. ML researchers have recently recognized the appeal of psychometric and other psychology-inspired techniques for measuring performance (Hérnandez-Orallo et al., 2017). Item response theory (IRT) provides a superior method to assess individual-level performance on decision-related tasks that goes beyond simply averaging performance over test items (Wilmer et al., 2012). Instead, IRT measures performance as a function of item-related difficulty and respondent-related ability (Lord, 2012). Recognizing this duality is important because all ML ability is a function of the data it was trained on. Because IRT can be used to estimate the item-level difficulties and respondent ability separately, it is an excellent candidate method for comparing individual differences in AI, especially for classifiers, where latent variables guide decision-making (Martínez-Plumed et al., 2019). When known variables guide decision-making, evidenced by varying performance as a function of stimulus intensity, classic psychophysics methods can be used to provide detailed descriptions of ability. RichardWebster et al. (2019) used psychophysical methods to characterize model performance as a function of stimulus degradation. The result was a family of performance-based curves that made comparing models easy and intuitive. Finally, researchers should consider using signal detection theory (SDT) to model individual-level performance when the AI is required to make decisions under uncertainty (Macmillan & Creelman, 2004). Originally used to model radar monitoring—a noisy visual perception task—SDT gives us a tool to specify perceptual sensitivity and decision-related biases by measuring the distribution of hits and false alarms (Witt et al., 2015).

Practical & ethical considerations in the laboratory

Human subjects research, and to a lesser extent non-human animal research, is regulated by extensive ethical mandates enforced by internal review boards at universities and other research institutions. These review boards are empowered by governments to ensure the safety, well-being, and privacy of human subjects, in response to psychology and medicine's histories of unethical practices (General Assembly of the World Medical Association, 2014). Machines, on the other hand, currently have no rights, and IRBs that regulate machine learning research are rare (Metcalf & Crawford, 2016). This is not to say that there are no ethical guidelines in machine learning. A growing field is dedicating itself to the ethical treatment and collection of data, the fairness and debiasing of the data, and the way machines are deployed. But there are no limits to what can be done to an algorithm for the purposes of research.

Another consideration that will differ for psychology research on machines versus humans is the rate of data acquisition. A majority of human subjects research in cognitive psychology is carried out on undergraduate psychology students (Rad et al., 2018; Henrich et al., 2010), who participate in exchange for course credit or money. This subject pool is a limited, time-sensitive resource that anecdotally peaks at the start of the Fall semester, when students are racing to complete their requirements (n.b. this manuscript was written pre-COVID, which has changed the way psychologists collect data, with many moving entirely online. It is too early to tell whether this will increase or decrease the rate of human data acquisition and what impact it will have on the timelines of research projects). If the reader generously assumes that training and debriefing a participant requires 10 min, and that a trial of an experimental protocol takes 10 s, a researcher can gather only 300 observations from any given participant, introducing a danger of underpowering estimates of central tendency and variability or restricting the number of cells in the experimental design. This problem must seem absurd to a machine learning researcher, who can collect massive data sets from any given subject. In comparison, OpenAI's hide-and-seek RL agents' behavior evolved over 500 million test trials (Baker et al., 2019); literally a tireless, sleepless human lifetime of laboratory-equivalent behavior. All of this

is to say that research in Artificial Cognition has a different set of operational bottlenecks compared to human cognitive psychology. There is no time-limited crush of intro-psych students getting research credits in September, there is no summer lull, and there is no need to spend grant money on paid participants. There is no protracted IRB review, and no tedious data storage and management. Machine Learning researchers face a different set of bottlenecks: access to rich datasets, which is always accelerating, and where training AI is part of the experiment, access to computing resources, which can mean time, money, and/or access to dedicated clusters. Often, experiments in Artificial Cognition will not require training anything new—for example when experimenting on existing, parameterized models. In these cases, conducting psychology experiments on machines will be limited only by the researcher's imagination and how quickly they can design and deploy a new experiment.

Conclusions

This paper was written as an open-ended research proposal with the goal of motivating other psychologists to participate in this new field. We have described the similarities between the black box challenge in deep learning and human cognition. Consumers, legislators, and machine learning researchers are interested in new approaches to XAI, and we think that cognitive psychologists have the tools and the tradition to produce satisfying models of behavior. To that end, we have provided justification for a cognition-inspired approach to XAI, we have reviewed the XAI field with non-computer scientists as a target audience, and we have suggested a framework for performing experiments on machines with an accompanying tutorial noting critical differences versus experimenting on humans. We hope that we have convinced the reader that artificial cognition presents a unique opportunity for cognitive psychologists to engage in a rigorous academic pursuit in an applied setting with an urgent need.

Open Practices Statement

There is no empirical work reported herein, but the two tutorials are open and accessible at the links provided in "Getting started".

Acknowledgements Thanks to Matthew Hilchey, Magdalena Sobol, Angus Galloway, Michael Ridley, and Bart Gajderowicz for helpful comments on an earlier version of this manuscript. Thanks to Shashank Shekhar for contributing code to the tutorial. Thanks to Chaz Firestone and two anonymous reviewers for helpful critiques.

This research was developed with funding from DARPA. The views, opinions and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. government. The authors also acknowledge support from the Canadian Institute for Advanced Research and the Canada Foundation for Innovation. Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute.

References

Adebayo, J., & Kagal, L. (2016). Iterative Orthogonal, Feature Projection for Diagnosing Bias in Black-Box Models. arXiv:1611.04967 [cs, stat].

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity Checks for Saliency Maps. *Advances in Neural Information Processing Systems*, 11.

Ancona, M., Ceolini, E., Öztireli, C., & Gross, M. (2018). Towards better understanding of gradient-based attribution methods for Deep, Neural Networks. arXiv:1711.06104 [cs, stat].

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., ..., Rahwan, I. (2018). The Moral Machine experiment. *Nature*, *563*(7729), 59–64.

Baker, B., Kanitscheider, I., Markov, T., Wu, Y., Powell, G., McGrew, B., & Mordatch, I. (2019). Emergent Tool, Use From Multi-Agent Autocurricula. arXiv:1909.07528 [cs, stat].

Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gutfreund, D., ..., Katz, B. (2019). Objectnet: a large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems*, pp. 9453–9463.

Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness in Machine Learning - 2019-09-19. *Fordham Law Review*, 28.

Bayat, A., Do Koh, H., Kumar Nand, A., Pereira, M., & Pomplun, M. (2018). Scene grammar in human and machine recognition of objects and scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1992–1999.

Bellemare, M. G., Naddaf, Y., Veness, J., & Bowling, M. (2013). The arcade learning environment: an evaluation platform for general agents. *Journal of Artificial Intelligence Research*, *47*, 253–279.

Bien, J., & Tibshirani, R. (2011). Prototype selection for interpretable classification. *The Annals of Applied Statistics*, 2403–2424.

Bjork, R. A. (1989). An adaptive mechanism in human memory. *Varieties of memory and consciousness: Essays in honour of Endel Tulving*, 309–330.

Bojarski, M., Yeres, P., Choromanska, A., Choromanski, K., Firner, B., Jackel, L., & Muller, U. (2017). Explaining How, a Deep Neural Network Trained with End-to-End Learning Steers a Car. arXiv:1704.07911 [cs].

Brendel, W., & Bethge, M. (2019). Approximating CNNs with bag-of-local-features models works surprisingly well on ImageNet. arXiv:1904.00760

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pp. 77–91.

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-Dependent Algorithm Aversion. *Journal of Marketing Research*, *56*(5), 809–825.

Cheung, B., Weiss, E., & Olshausen, B. (2016). Emergence of foveal image sampling from learning to attend in visual scenes. arXiv:1611.09430 [cs].

Chomsky, N. (1957). Syntactic structures. The Hague: Mouton.. 1965. Aspects of the theory of syntax. *Cambridge, Mass.: MIT Press.(1981) Lectures on Government and Binding, Dordrecht: Foris.(1982) Some Concepts and Consequences of the Theory of Government and Binding. LI Monographs*, *6*, 1–52.

Chomsky, N. (1959). A review of bf skinner's verbal behavior. *Language*, *35*(1), 26–58.

de Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A. B., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied*, *22*(3), 331–349.

Deisenroth, M. P., Faisal, A. A., & Ong, C. S. (2020). *Mathematics for Machine Learning*.

Despraz, J., Gomez, S., Satizábal, H. F., & Pena-reyes, C. A. (2017). Towards a Better Understanding of Deep Neural Networks Representations using Deep Generative Networks. *Proceedings of the 9th International Joint Conference on Computational intelligence*, pp. 215–222, Funchal, Madeira, Portugal. SCITEPRESS - Science and Technology Publications.

Donders, F. C. (1868). Die schnelligkeit psychischer processe: Erster artikel. Archiv für Anatomie. *Physiologie und wissenschaftliche Medicin*, 657–681.

Edwards, L., & Veale, M. (2018). Enslaving the Algorithm: From a "Right to an Explanation" to a "Right to Better Decisions"? *IEEE Security & Privacy*, *16*(3), 46–54.

Elsayed, G., Shankar, S., Cheung, B., Papernot, N., Kurakin, A., Goodfellow, I., & Sohl-Dickstein, J. (2018). Adversarial examples that fool both computer vision and time-limited humans. In *Advances in Neural Information Processing Systems*, pp. 3910–3920.

Erhan, D., Bengio, Y., Courville, A., Vincent, P., & Box, P. O. (2009). Visualizing Higher-Layer Features of a Deep Network. Technical report, University of Montreal.

Fisher, A., Rudin, C., & Dominici, F. (2018). Model class reliance: Variable importance measures for any machine learning model class, from the "Rashomon" perspective. arXiv:1801.01489

Fong, R., & Vedaldi, A. (2017). Interpretable Explanations, of Black Boxes by Meaningful Perturbation. 2017 IEEE International Conference on Computer Vision (ICCV), pages 3449–3457. arXiv:1704.03296

Frank, D.-A., Chrysochou, P., Mitkidis, P., & Ariely, D. (2019). Human decision-making biases in the moral dilemmas of autonomous vehicles. *Scientific Reports*, *9*(1), 13080.

Frankle, J., & Carbin, M. (2019). The Lottery, Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. arXiv:1803.03635 [cs].

Frosst, N., & Hinton, G. (2017). Distilling a Neural, Network Into a Soft Decision Tree. arXiv:1711.09784 [cs, stat].

Geirhos, R., Temme, C. R., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. In *Advances in neural information processing systems*, pp. 7538–7550.

General Assembly of the World Medical Association (2014). World Medical Association Declaration of Helsinki: Ethical principles for medical research involving human subjects. *The Journal of the American College of Dentists*, *81*(3), 14.

Ghorbani, A., Wexler, J., Zou, J. Y., & Kim, B. (2019). Towards automatic concept-based explanations. In *Advances in Neural Information Processing Systems*, pp. 9277–9286.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining Explanations: An Overview of Interpretability of Machine Learning. arXiv:1806.00069 [cs, stat].

Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, *24*(1), 44–65.

Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, *38*(3), 50–57. arXiv:1606.08813

Gu, T., Liu, K., Dolan-Gavitt, B., & Garg, S. (2019). BadNets: Evaluating Backdooring Attacks on Deep Neural Networks. *IEEE Access*, *7*, 47230–47244.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*, *51*(5), 1–42.

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. arXiv:1803.11138

Gunning, D., & Aha, D. W. (2019). DARPA's Explainable Artificial Intelligence Program. p. 16.

Gunning, D. (2017). Explainable Artificial Intelligence (XAI) - DARPA.

Henderson, M. M., & Serences, J. (2020). Biased orientation representations can be explained by experience with non-uniform training set statistics. *bioRxiv*.

Hendricks, L. A., Hu, R., Darrell, T., & Akata, Z. (2018). Grounding visual explanations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 264–279.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2-3), 61–83.

Hinton, G. E. (1986). Learning distributed representations of concepts. In *Proceedings of the Eighth Annual Conference of the Cognitive Science Society*, pp. 1–12.

Hoffman, R., Miller, T., Mueller, S. T., Klein, G., & Clancey, W. J. (2018a). Explaining explanation, Part 4: A Deep Dive on Deep Nets. *IEEE Intelligent Systems*, *33*(3), 87–95.

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018b). Metrics for explainable AI: Challenges and prospects. arXiv:1812.04608

Hérnandez-Orallo, J., Baroni, M., Bieger, J., Chmait, N., Dowe, D. L., Hofmann, K., . . . , Thórisson, K. R. (2017). A New, AI Evaluation Cosmos: Ready to Play the Game?. *AI Magazine*, *38*(3), 66–69.

Huang, G., Chen, D., Li, T., Wu, F., van der Maaten, L., & Weinberger, K. Q. (2017). Multi-scale dense networks for resource efficient image classification. arXiv:1703.09844

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., & Madry, A. (2019). Adversarial Examples, Are Not Bugs, They Are Features. arXiv:1905.02175 [cs, stat].

Jain, S., & Wallace, B. C. (2019). Attention is not explanation. arXiv:1902.10186

Kalra, N., & Paddock, S. M. (2016). Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability? *Transportation Research Part A: Policy and Practice*, *94*, 182–193.

Kim, B., Khanna, R., & Koyejo, O. O. (2016). Examples are not enough, learn to criticize! criticism for interpretability. In *Advances in neural information processing systems*, pp. 2280–2288.

Kim, J., & Canny, J. (2017). Interpretable learning for self-driving cars by visualizing causal attention. In *Proceedings of the IEEE international conference on computer vision*, pp. 2942–2950.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (2018). Interpretability Beyond, Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). arXiv:1711.11279 [stat].

Kim, B., Reif, E., Wattenberg, M., & Bengio, S. (2019). Do neural networks show gestalt phenomena? an exploration of the law of closure. arXiv:1903.01069

Kingstone, A., Smilek, D., & Eastwood, J. D. (2008). Cognitive ethology: a new approach for studying human cognition. *British Journal of Psychology*, *99*(3), 317–340.

Klein, G. (2018). Explaining explanation, Part 3: The Causal Landscape. *IEEE Intelligent Systems*, *33*(2), 83–88.

Krause, J., Perer, A., & Ng, K. (2016). Interacting with Predictions: Visual Inspection of Black-box Machine Learning Models. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*, (pp. 5686–5697). Santa Clara: ACM Press.

Krening, S., Harrison, B., Feigh, K. M., Isbell, C. L., Riedl, M., & Thomaz, A. (2016). Learning from explanations using sentiment and advice in RL. *IEEE Transactions on Cognitive and Developmental Systems*, *9*(1), 44–55.

Kuang, C., & Can, A. I. (2017). Be Taught to Explain Itself?*The New York Times*, 7.

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development*, *3*(3), 299–321.

Landau, B., Smith, L. B., & Jones, S. (1992). Syntactic context and the shape bias in children's and adults' lexical learning. *Journal of Memory and Language*, *31*(6), 807–825.

Lee, J., Shin, J.-H., & Kim, J.-S. (2017). Interactive visualization and manipulation of attention-based neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 121–126).

Leibo, J. Z., d'Autume, C. d. M., Zoran, D., Amos, D., Beattie, C., Anderson, K., ..., Botvinick, M. M. (2018). Psychlab: A, Psychology Laboratory for Deep Reinforcement Learning Agents. arXiv:1801.08116 [cs, q-bio].

Lillicrap, T. P., & Kording, K. P. (2019). What does it mean to understand a neural network? arXiv:1907.06374 [cs, q-bio, stat].

Lipton, Z. C. (2017). The Mythos, of Model Interpretability. arXiv:1606.03490 [cs, stat].

Lord, F. M. (2012). *Applications of item response theory to practical testing problems*: Routledge.

Maaten, L. v. d., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605.

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide*: Psychology Press.

Madumal, P., Miller, T., Sonenberg, L., & Vetere, F. (2019). Explainable Reinforcement, Learning Through a Causal Lens. arXiv:1905.10958 [cs, stat].

Martínez-Plumed, F., Prudêncio, R. B., Martínez-usó, A., & Hernández-orallo, J. (2019). Item response theory in AI,: Analysing machine learning classifiers at the instance level. *Artificial Intelligence*, *271*, 18–42.

Metcalf, J., & Crawford, K. (2016). Where are human subjects in big data research? the emerging ethics divide. *Big Data & Society*, *3*(1), 2053951716650211.

Miller, T. (2017a). Explanation in Artificial, Intelligence: Insights from the Social Sciences. arXiv:1706.07269 [cs].

Miller, T., Howe, P., & Sonenberg, L. (2017b). Explainable AI,: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences. arXiv:1712.00547 [cs].

Mnih, V., Heess, N., & Graves, A. (2014). Recurrent models of visual attention. In *Advances in neural information processing systems* (pp. 2204–2212).

Molnar, C. (2019). Interpretable Machine Learning. https://christophm.github.io/interpretable-ml-book/

Mott, A., Zoran, D., Chrzanowski, M., Wierstra, D., & Rezende, D. J. (2019). Towards Interpretable, Reinforcement Learning Using Attention Augmented Agents. arXiv:1906.02500 [cs, stat].

Mueller, S. T., Hoffman, R. R., Clancey, W., & Emrey, A. (2019). Explanation in Human-AI Systems: A Literature Meta-Review Synopsis of Key Ideas and Publications and Bibliography for Explainable AI. Technical report, Institute for Human and Machine Cognition.

Nairne, J. S. (2011). Adaptive memory: Nature's criterion and the functionalist agenda. *The American Journal of Psychology*, *124*(4), 381–390.

Navon, D. (2003). What does a compound letter tell the psychologist's mind? *Acta Psychologica*, *114*(3), 273–309.

Neisser, U. (1978). Memory: What are the important questions. *Memory observed: Remembering in natural contexts*, 3–19.

Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 427–436). Boston: IEEE.

Nguyen, A., Dosovitskiy, A., Yosinski, J., Brox, T., & Clune, J. (2016). Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. *Advances in neural information processing systems*, 3387–3395.

Norvig, P. (2017). Google's approach to artificial intelligence and machine learning — a conversation with peter norvig.

of the European Union, C. (2016). Regulation (EU) 2016/ 679 of The European Parliament and of The Council - of 27 April 2016 - on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/ 46/ EC (General Data Protection Regulation).

Öhlschläger, S., & Võ, M. L.-H. (2017). SCEGRAM: An image database for semantic and syntactic inconsistencies in scenes. *Behavior Research Methods*, *49*(5), 1780–1791.

Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., & Mordvintsev, A. (2018). The building blocks of interpretability. Distill, https://distill.pub/2018/building-blocks

Osband, I., Doron, Y., Hessel, M., Aslanides, J., Sezener, E., Saraiva, A., ..., Van Hasselt, H. (2019). Behaviour Suite, for Reinforcement Learning. arXiv:1908.03568 [cs, stat].

Papernot, N., & McDaniel, P. (2018). Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. arXiv:1803.04765

Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, *62*(3), 54–60.

Popper, K. (2014). *Conjectures and refutations: The growth of scientific knowledge*. Routledge.

Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, *32*(1), 3–25.

Rad, M. S., Martingano, A. J., & Ginges, J. (2018). Toward a psychology of homo sapiens: Making psychological science more representative of the human population. *Proceedings of the National Academy of Sciences*, *115*(45), 11401–11405.

Rahwan, I., & Cebrian, M. (2018). Machine Behavior Needs to Be an Academic Discipline. *Nautilus*, 8.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., ..., Wellman, M. (2019). Machine Behaviour. *Nature*, *568*(7753), 477–486.

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (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. *The Journal of Neuroscience*, *38*(33), 7255–7269.

Ribeiro, M. T., Singh, s., & Guestrin, C. (2016). "Why, Should I Trust You?": Explaining the Predictions of Any Classifier. arXiv:1602.04938 [cs, stat].

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High Precision Model-Agnostic Explanations. pp. 9.

RichardWebster, B., Yon Kwon, S., Clarizio, C., Anthony, S. E., & Scheirer, W. J. (2018). Visual psychophysics for making face recognition algorithms more explainable. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 252–270).

RichardWebster, B., Anthony, S. E., & Scheirer, W. J. (2019). PsyPhy: A Psychophysics Driven Evaluation Framework for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(9), 2280–2286. arXiv:1611.06448

Ritter, S., Barrett, D. G., Santoro, A., & Botvinick, M. M. (2017). Cognitive psychology for deep neural networks: a shape bias case study. In *International Conference on Machine Learning* (2940–2949).

Scheirer, W. J., Anthony, S. E., Nakayama, K., & Cox, D. D. (2014). Perceptual Annotation: Measuring Human Vision to Improve Computer Vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(8), 1679–1686.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 618–626).

Serrano, S., & Smith, N. A. (2019). Is attention interpretable? arXiv:1906.03731

Sheh, R., & Monteath, I. (2018). Defining Explainable AI for Requirements Analysis. *KI - Künstliche Intelligenz*, *32*(4), 261–266.

Si, Z., & Zhu, S.-C. (2013). Learning and-or Templates for Object Recognition and Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(9), 2189–2205.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv:1312.6034

Sloman, S. (2005). *Causal models: How people think about the world and its alternatives*: Oxford University Press.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. arXiv:1312.6199 [cs].

Taylor, E., Shekhar, S., & Taylor, G. W. (2020). Response time analysis for explainability of visual processing in CNNs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 382–383).

Tinbergen, N. (1963). On aims and methods of ethology. *Zeitschrift Für Tierpsychologie*, *20*(4), 410–433.

Van Looveren, A., & Klaise, J. (2019). Interpretable Counterfactual, Explanations Guided by Prototypes. arXiv:1907.02584 [cs, stat].

S

Vinyals, O., Ewalds, T., Bartunov, S., Georgiev, P., Vezhnevets, A. S., Yeo, M., . . . , Tsing, R. (2017). StarCraft, II: A New Challenge for Reinforcement Learning. arXiv:1708.04782 [cs].

Võ, M. L.-H., & Wolfe, J. M. (2013). Differential electrophysiological signatures of semantic and syntactic scene processing. *Psychological Science*, *24*(9), 1816–1823.

Volokitin, A., Roig, G., & Poggio, T. A. (2017). Do deep neural networks suffer from crowding?. In *Advances in Neural Information Processing Systems* (pp. 5628–5638).

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, *12*(3), 129–140.

Wertheimer, M. (1923). Laws of organization in perceptual forms. *A source book of Gestalt Psychology*.

Whitney, D., & Levi, D. M. (2011). Visual crowding: a fundamental limit on conscious perception and object recognition. *Trends in Cognitive Sciences*, *15*(4), 160–168.

Wilmer, J. B., Germine, L., Chabris, C. F., Chatterjee, G., Gerbasi, M., & Nakayama, K. (2012). Capturing specific abilities as a window into human individuality: The example of face recognition. *Cognitive Neuropsychology*, *29*(5-6), 360–392.

Witt, J. K., Taylor, J. E. T., Sugovic, M., & Wixted, J. T. (2015). Signal detection measures cannot distinguish perceptual biases from response biases. *Perception*, *44*(3), 289–300.

Wolfe, J. M., & Gray, W. (2007). Guided search 4.0. *Integrated models of cognitive systems*, 99–119.

Wolfe, J. M. (2016). Rethinking the basic-applied dichotomy.

Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., & Lipson, H. (2015). Understanding Neural, Networks Through Deep Visualization. arXiv:1506.06579 [cs].

Zeiler, M. D., & Fergus, R. (2013). Visualizing and Understanding, Convolutional Networks. arXiv:1311.2901 [cs].

Zhang, Q., Wu, Y. N., & Zhu, S.-C. (2018). Interpretable Convolutional Neural Networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (pp. 8827–8836). Salt Lake City: IEEE.

Zhou, Z., & Firestone, C. (2019). Humans can decipher adversarial images. *Nature Communications*, *10*(1), 1–9.

Zilke, J. R., Mencía, E. L., & Janssen, F. (2016). DeepRED–rule extraction from deep neural networks. In *International Conference on Discovery Science*, (pp. 457–473): Springer.