Does Explainable AI Need Cognitive Models?

Céline Budding (c.e.budding@tue.nl)
Philosophy & Ethics, Eindhoven University of Technology
Eindhoven, The Netherlands

Carlos Zednik (c.a.zednik@tue.nl)
Philosophy & Ethics, Eindhoven University of Technology
Eindhoven, The Netherlands

Abstract

Explainable AI (XAI) aims to explain the behavior of opaque AI systems, and in this way, increase their trustworthiness. However, current XAI methods are explanatorily deficient. On the one hand, "top-down" XAI methods allow for global and local prediction, but rarely support targeted internal interventions. On the other hand, "bottom-up" XAI methods may support such interventions, but rarely provide global behavioral predictions. To overcome this limitation, we argue that XAI should follow the lead of cognitive science in developing *cognitive models* that simultaneously reproduce behavior and support interventions. Indeed, novel methods such as *mechanistic interpretability* and *causal abstraction analysis* already reflect cognitive modeling principles that are familiar from the explanation of animal and human intelligence. As these methods might serve as best practices for trustworthy AI, they deserve closer philosophical scrutiny.

Keywords: Explainable AI; cognitive models; trustworthiness; interventions

Introduction

Many recent milestones in Artificial Intelligence (AI) have been achieved through the use of machine learning. Unfortunately, many state-of-the-art AI systems are *opaque*: it is difficult to know what they do, why they do what they do, and how they work (Zednik, 2021). Among other drawbacks, opacity undermines *trustworthiness*, which according to EU AI Ethics guidelines requires that humans be able to accurately describe and control AI systems' behaviors (High-Level Expert Group on AI, 2019). For example, banks seeking to comply with principles of anti-discrimination are unlikely to trust an algorithmic credit scoring application if they cannot predict that it will treat women equally to men. Similarly, vehicle manufacturers should not trust a vision module for autonomous driving if they do not know how to intervene on that module if it fails to reliably detect cyclists.

Where opacity is a problem, explainable AI (XAI) is often considered a solution. Indeed, many different XAI methods have been developed to answer questions about what AI systems actually do, why they do it, and how they work. Nevertheless, a closer look at the methodological principles governing these methods reveals that they are likely to fall short of ensuring trustworthiness. Indeed, XAI methods traditionally follow either one of two explanatory strategies, each of which can be likened to longstanding research strategies in the study of human and animal behavior, and each of which has characteristic advantages and disadvantages. On the one hand, "top-down" XAI methods that follow the basic logic of psychological explanation (see e.g. Rahwan et al., 2019; Taylor & Taylor, 2021) may adequately describe an AI system's overt behavior, but typically fall short of providing insights into the underlying processes or mechanisms that can be targeted by systematic interventions. On the other hand, "bottom-up" methods that adopt the logic of neuroscientific investigation (see e.g. Olah, Mordvintsev, & Schubert, 2017; Lam, 2022) focus on describing causal and computational structures at the level of layers, nodes, or learned representations, but mostly fail to account for AI systems' global behavior in real-world contexts. Insofar as trust in AI technology is grounded in an ability to describe and control its behavior, traditional XAI methods provide insufficient grounds for trust.

In this contribution, we advance the idea that, in order to adequately promote trustworthiness, explainable AI should adopt a dually constrained explanatory strategy familiar from cognitive science. This strategy centers on the development of *cognitive models* (Busemeyer & Diederich, 2010): abstract mathematical or computational representations of the cognitive processes that are realized in a system's physical hardware and that are causally responsible for that system's behavior. As such, cognitive models are constrained by behavior "from above" and by causal structures "from below", and are for this reason able to simultaneously reproduce behavior and to support systematic and targeted interventions. While current XAI methods typically either reproduce behavior or support interventions, explanations that center on cognitive models are capable of doing both.

More specifically, the goals of this contribution are threefold. First, we argue that current XAI methods insufficiently promote trustworthiness because they are either "top-down" or "bottom-up". Second, we argue that a unified explanatory strategy grounded in cognitive modeling may be better suited to promoting trustworthiness because it simultaneously allows for reproduction of behavior and support for intervention. Finally, we introduce two very recent XAI methods—*mechanistic interpretability* (Elhage et al., 2021) and *causal abstraction analysis* (Geiger, Lu, Icard, & Potts, 2021)—and argue that these methods may be promising first examples of cognitive modeling in explainable AI.

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

State-of-the-art XAI is explanatorily deficient

Explainable AI aims to explain the behavior of opaque AI systems using mathematical and computational methods. In this section, we will classify XAI methods into two broad groups, distinguished by their basic logic or strategy: XAI methods that follow a "top-down" explanatory strategy that broadly resembles methods from cognitive psychology, versus those that follow a "bottom-up" explanatory strategy that follows the logic of neuroscientific investigation. While it is not possible to provide a comprehensive review of XAI methods, we will provide an illustrative example for each group and identify the general advantages and disadvantages of each.

Artificial Psychology: "Top-down" XAI

One way to explain an AI system's behavior is to take a "top-down" approach that bears methodological resemblance to the basic logic of cognitive psychology (for discussion, see Rahwan et al., 2019; Taylor & Taylor, 2021). The starting point of this approach is to observe a particular AI system's local behavior in a specific experimental setting, or its global behavior over a wide range of contexts. For instance, one could study a visual classifier's local behavior by describing the effect of perturbations on its classification performance, or its global behavior for an entire dataset by keeping track of those instances for which its classification is incorrect. Then, in a direct parallel to many scientific investigations of human and animal behavior, a mathematical or computational model is constructed to adequately reproduce the observed behavior (description) and predict future behavior (prediction).

One well-known "top-down" XAI method is *Local Interpretable Model-Agnostic Explanation* (LIME, Ribeiro, Singh, & Guestrin, 2016). LIME aims to determine the features of an input that have a particularly high effect on the system's output. That is, the goal is to determine why the system returned a particular output by analyzing which input features would have changed its prediction. While the decision boundary of the original system is likely to be nonlinear, the assumption is that it can be locally approximated as a linear function near a single data point. As this linear approximation is much simpler than the original nonlinear decision boundary, it can be used to infer how the output would have been different for small perturbations in the input. In this way, LIME adequately describes local system behavior by constructing a simpler *surrogate model* that reproduces the important aspects of the AI system's behavior.

While LIME reproduces a system's local behavior for individual inputs, other "top-down" XAI methods approximate global behavior for all inputs. For example, *tree-extraction methods* (Wu et al., 2018) use a system's global input/output behavior to train a decision tree that adequately reproduces this behavior. Insofar as the decision tree is likely to be smaller and simpler than the original system–e.g. a deep neural network–this method can be used to develop a surrogate model that approximates the original system's global behavior to an arbitrary degree of precision, but that requires less computational resources to run.

LIME, tree-extraction, and other "top-down" XAI methods describe and predict AI systems' behaviors by constructing simplified models that reproduce these systems' observed behaviors and predict unobserved behaviors. In cognitive psychology, analogous "top-down" methods are traditionally deployed because the underlying processes or mechanisms are unknown and cannot easily be studied. Relatedly, in artificial intelligence these methods are helpful because they can be used to describe the behavior of systems whose implementation is particularly complex or unavailable to the public (Burrell, 2016).

Despite these advantages, "top-down" XAI methods are explanatorily deficient in light of the demands posed by trustworthiness. While these methods may succeed at describing and predicting a system's behavior, there is no guarantee that the internal structure of the model actually reflects the causally-relevant features of the original system. While methods like LIME and tree-extraction might provide plausible "explanatory stories" that fit the behavioral data (e.g. by indicating that the presence of pointy ears is important for classifying an image as "cat"), there is no guarantee that these stories are actually true (e.g. the system internally represents the presence of pointy ears and deploys that representation to classify an image as a cat, as opposed to doing something else). As such, while "top-down" methods successfully predict and reproduce behavior, they do not typically allow for the application of surgical interventions on the system's internal elements. That is, "top-down" XAI methods do not typically identify causally-relevant structures within the system–such as nodes or other representational structures–that could be targeted by interventions so as to change the system's behavior.

Artificial Neuroscience: "Bottom-up" XAI

While "top-down" XAI methods take an AI system's behavior as the starting point, "bottom-up" XAI methods focus on the processing within the system. This general approach is familiar from neuroscience, in which various methods such as single-cell recording or fMRI imaging are used to identify and describe the neural structures and functions that contribute to a particular behavioral or cognitive capacity (Bechtel, 2007). Similarly, "bottom-up" XAI methods study particular elements of the AI system (e.g. a network's unit activations or connection weights) and try to characterize how these components contribute to the behavior of the system as a whole. Similar to the way in which neuroscience might investigate the way in which the spike pattern of a particular neuron might contribute to e.g. visual perception, "bottom-up" XAI methods determine which role particular elements of the network play in its overt behavior (for discussion, see e.g. Lam, 2022).

A notable example of "bottom-up" XAI is *feature optimization* (Olah et al., 2017). Feature optimization focuses on a particular element of an AI system, such as an individ-

ual node or layer, and tries to determine what kind of input this node is particularly sensitive to or activated by. This is generally done by optimizing synthetic inputs, which are more likely to lead to maximal activation than actual inputs from the training set. In this way, one can develop a set of images—in the case of a vision system—that allow us to infer the kinds of input features the element in question "detects". Thus for example, Olah and colleagues identified nodes that were sensitive to curves, and even so-called circuits–groups of nodes–that detect dog heads.

The main advantage of "bottom-up" methods like feature optimization is that they are frequently able to identify particular structures within a network that can be thought to detect or represent recognizable features such as edges, ears, or balls. As the focus is on actual structures within the network, these methods could support targeted interventions: if a network appears to be using a particular unit or layer to detect a particular feature or object, it might be sufficient to suppress or promote the relevant unit activations to change the system's overall ability to classify certain kinds of images in a predictable way.

Nevertheless, "bottom-up" methods like feature optimization also fall short of satisfying the requirements for trustworthiness. One issue is that synthetic inputs generally have limited naturalistic value and can be hard to interpret using an everyday conceptual repertoire (Bau, Zhou, Khosla, Oliva, & Torralba, 2017; Borowski et al., 2020). Moreover, feature optimization generally focuses on single nodes or small groups of nodes. While these structures can be thought to detect particular input features, feature optimization does not provide insight into how these nodes interact with other nodes in the system to drive the network's overall behavior. That is, while feature optimization might suggest that activity in a particular node correlates with the presence of pointy ears in cats, this does not guarantee that this particular node actually plays a causal role in the prediction of "cat". As such, "bottom-up" methods like feature optimization fall short of predicting global system behavior; at most, these methods can be used to describe and influence a system's responses to the presence or absence of very specific features. While these methods therefore support interventions to a limited extent, it remains unclear how useful such limited interventions are for controlling AI systems' behaviors in complex and dynamic real-world environments.

Cognitive modeling as an explanatory strategy

As shown in the previous section, current XAI methods provide insufficient grounds for trust. On the one hand, "top-down" methods describe and predict behavior, but do not support surgical interventions that allow us to modify that behavior. On the other hand, "bottom-up" methods promise to support such interventions, but have limited ability to reproduce global behavior. Insofar as trust requires a combination of prediction and intervention, there is a need for better explanation.

In section 4 below, we will propose that explainable AI might take inspiration from one of the main explanatory strategies of cognitive science, centering on the development of *cognitive models*. Cognitive models are scientific representations of cognitive processes.[1] That is, they represent the "mental" structures and functions that contribute to a particular behavior or that underlie a particular cognitive capacity at an algorithmic level of analysis (Marr, 1982). Of course, there is much mystery, and thus little agreement, about what these structures and functions are actually like. Although there is no need to be overly specific, it will suffice to assume that cognitive processes are realized or implemented in neural "hardware", and that they can be likened to the "software" that governs the causal processes that eventually give rise to some observed behavior or capacity (figure 1).

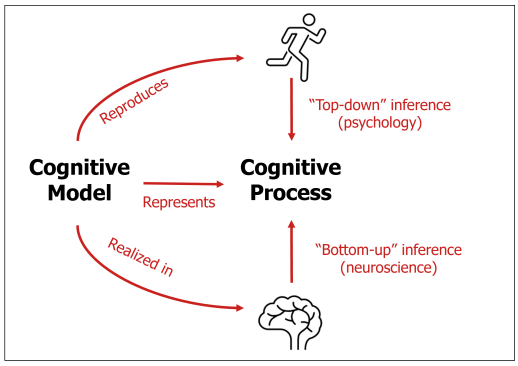

Figure 1: **A schematic overview of cognitive modeling.** Cognitive models represent cognitive processes. In doing so, they reproduce overt behavior, and are plausibly realized in physical hardware. Investigators may infer the structure of cognitive processes (and thus, decide on the structure of a cognitive model) from the "top down", by focusing on behavior, or from the "bottom up", by focusing on the hardware.

Most cognitive models take the form of mathematical or computational descriptions that represent the relevant cognitive processes (Busemeyer & Diederich, 2010). This kind of representation has numerous advantages. For one, it allows the model to be implemented on a digital computer, and for its behavior to be simulated so as to derive quantitative as well as qualitative predictions. For another, it allows investigators to easily intervene on the model, e.g. by modifying the value of a parameter, or by changing a specific architectural feature. Finally, it is worth noting that certain mathematical or computational descriptions may be considered more *interpretable* than others, in the sense of being comparatively low in size and complexity, and in the sense of containing structures or functions that are meaningful to a human investigator. Al-

[1]This notion of 'cognitive model' in the sense of a model that is used by scientists to explain cognitive or behavioral phenomena is not to be confused with the notion of a 'cognitive model' (also often called a 'mental model') in the sense of a mental representation that drives a particular cognitive system's behavior.

though there is no general requirement that a cognitive model be interpretable in this sense—indeed, many of the best cognitive models, including many neural network models, are not interpretable—interpretability does offer specific advantages during model development and model operation.

Although the development of cognitive models remains somewhat of an ill-understood "dark art" (Zednik & Jäkel, 2016), there are a number of rather general model-development strategies that are worth highlighting in part because they may also be applied in the context of explainable AI. One common strategy, when confronted with the challenge of explaining a particular behavior or capacity, is to consider pre-existing engineering solutions for tasks that resemble the behavior being explained. This strategy, which Gerd Gigerenzer (1991) has previously called *tools-to-theories*, involves postulating that an algorithm that has been successfully deployed in engineering applications (e.g. Monte Carlo sampling for estimating probability distributions, or the drift-diffusion method for decision-making) is also a good description of a cognitive process. This strategy culminates in efforts to identify traces of the algorithm's activity in neural "hardware".

Another common model-development strategy is a divide-and-conquer strategy which Robert Cummins (1985) has previously termed *functional analysis*: analyzing a complex capacity (such as language-learning) into simpler capacities (such as learning of grammatical rules on the one hand and learning of semantic relations on the other), and then repeating the analytic process for each one of the simpler capacities. The strategy completes when simple capacities can be identified with the behavior of particular parts of a physical system (e.g., neurons or columns in the brain). If such identification succeeds, functional analysis contributes to what is commonly referred to as mechanistic explanation (Bechtel, 2007; Craver, 2007).

No matter how cognitive models are developed, the critical feature is that they are dually constrained. Whereas the "top-down" and "bottom-up" XAI methods that represent the current state-of-the-art focus on either behavioral data or the underlying system, good cognitive models always take into account both simultaneously. As a consequence, they are not only capable of reproducing behavior (e.g. by way of simulation on a digital computer), but also support systematic and surgical interventions (e.g. by way of modifying a model component or parameter). By combining "top-down" and "bottom-up" constraints, cognitive models can simultaneously ensure the reproduction of behavior and the support for intervention.

Cognitive models of artificial intelligence?

Given that cognitive models have been successfully deployed in cognitive science to explain the behavior of humans and animals, the question is whether they might also be useful for explaining the behavior of opaque AI systems. Although there is no need to assume that artificial intelligence closely resembles human or animal intelligence, it is apparent that the explanatory task facing XAI is closely analogous to the one facing cognitive science: explaining the behavior of high-dimensional nonlinear systems that interact with, learn from, and adapt to dynamic environments. Given that the task is similar, it is natural to consider the possibility that the solution might be similar as well. Specifically, it is worth considering whether the explanations being delivered in explainable AI can and should take the form of cognitive models.

In this section, we suggest that recent work on large language models (LLMs) has in fact begun to adopt a cognitive modeling approach. In particular, novel methods like *mechanistic interpretability* (MI, Elhage et al., 2021) and *causal abstraction analysis* (CAA, Geiger et al., 2021) use familiar model-development strategies to construct cognitive models that are constrained from both directions, and as such, reconstruct behavior while also supporting interventions. In contrast to more well-established XAI methods that deploy either a "top-down" or "bottom-up" approach in isolation, both mechanistic interpretability and causal abstraction analysis appear to do both simultaneously. Although it is unclear whether the adoption of cognitive modeling principles is intentional on the part of these methods' developers, inspecting them with these principles in mind can be helpful for better evaluating their potential to promote the development of trustworthy AI.

Mechanistic interpretability

Mechanistic interpretability (Elhage et al., 2021) is a novel approach that has been used to explain the language-learning ability of state-of-the-art LLMs. While the development of this method was inspired by results from feature optimization (Olah et al., 2017), it poses significant improvements in terms of facilitating trustworthiness. Specifically, MI aims to explain—in an interpretable way—the high-dimensional function an AI system learns during training by analyzing this problem into smaller component problems.

For example, to explain how LLMs perform next-word prediction, mechanistic interpretability divides the problem into smaller functional parts, such as in-context learning (figure 2). These smaller functional parts can then be identified with particular structures in the network. That is, MI aims to localize causally-relevant structures within the network and identify these with particular functional parts, i.e. aspects of the model's behavior. Notably, in this sense MI exemplifies the functional analysis strategy for cognitive model-development introduced above. Essential to the success of MI is the assumption that the internal processing of an AI system—consisting of e.g. its weights and activations—can be segmented into (potentially human-interpretable) variables. If this is indeed the case, particular behaviors of the network can be explained by describing these variables and the operations applied to them.

Elhage and colleagues take exactly this approach to investigate *in-context learning*—the system's ability to learn from examples—as one component of a large language model's

ability to perform next-word prediction. To identify the structures responsible for in-context learning, Elhage and colleagues first analyze a toy example: the smallest possible network that exhibits in-context learning. In this toy model, the authors identify so-called induction heads that play a role in performing the task. Specifically, when encountering a particular token *A*, induction heads "look back" into the text to find previous occurrences of A as well as the token that follows it, let's call it *B*. Induction heads then increase the likelihood of again predicting *B* in response to *A*. Finally, the authors show that such induction heads not only occur in the toy model, but also throughout the original LLM, thereby explaining the occurrence of in-context learning in the original system.

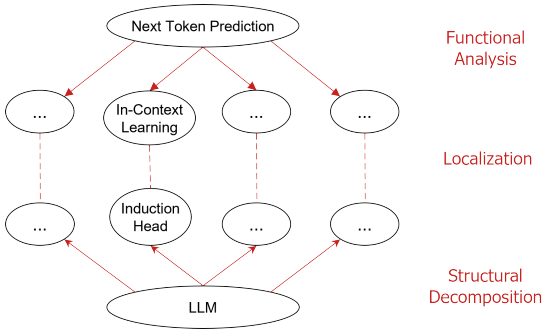

Figure 2: **A schematic overview of mechanistic interpretability.** First, a target behavior–such as next-token prediction–is analyzed into simpler component behaviors–such as in-context learning. Mechanistic interpretability then identifies particular structures in the network–in this case induction heads–in which the component behaviors can be localized. This instantiates the functional analysis model-development strategy widely invoked in cognitive science.

In this example, the explanation achieved through MI seems to both reproduce behavior and facilitate systematic interventions. First, the authors show that the toy model reproduces the target behavior—in-context learning—and identify the structure responsible for this behavior—induction heads. Then, they verify that the presence of induction heads is responsible for in-context learning in the original LLM, amongst others by "knocking out" induction heads, which is followed by a significant decrease in in-context learning. This illustrates that MI in fact identifies particular causally-relevant structures within the network, which can serve as targets for systematic interventions.

An advantage of MI is that, like many traditional "top-down" methods, it requires relatively little prior knowledge about the internal processing of the system, as the analysis and explanation start from the behavior of the target system and only then try to identify internal structures responsible for this behavior. Nevertheless, because it culminates in the identification of simple behavioral capacities localized in simple structural parts, the method is unlike traditional "top-down" methods in that it is likely to support systematic and targeted interventions on those parts.

Nevertheless, MI faces challenges as well. One challenge is that it is unclear to what extent functional parts of behavior can be localized to specific structures in the model internals. For example, it might be that many layers are involved in a particular aspect of behavior, in which case the explanation would likely be hard to interpret. Moreover, it remains unclear whether the identified structures provide appropriate targets for effective and systematic interventions. While intervention on e.g. induction heads might successfully change behavior, there is also a risk that this could lead to catastrophic interference.

Causal abstraction analysis

Another promising novel method is causal abstraction analysis (Beckers & Halpern, 2019; Geiger et al., 2021). Originally, CAA was developed to explain complex behaviors and phenomena that arise from systems that can be described at multiple levels of abstraction. For example, human behavior might be explained at the level of neuronal spike trains, or alternatively at the level of folk-psychological concepts such as beliefs and desires. At each of these levels, we can develop a causal model of how different variables interact, for example how a particular neuron affects the observed behavior. The question CAA aims to answer is whether a high-level causal model whose variables may for example represent "beliefs" and "desires" can be considered a faithful abstraction of a low-level causal model, such as one that represents neural activations (Beckers & Halpern, 2019).

More recently, CAA has been used specifically for XAI (Geiger et al., 2021). In this context, a trained neural network can be considered a low-level causal model. While this low-level network describes the causal relationships between variables—in this case nodes—these are generally hard to interpret. One way to explain such a low-level system could be to develop a high-level causal model that is a faithful abstraction of the original system, insofar as it contains the same causal relationships, but that is simpler and potentially more interpretable than the original system. Similar to MI, CAA thus aims to "break down" the internal processing of a system into high-level (and ideally, interpretable) causal variables that explain the observed behavior.

To this end, CAA starts with developing one or multiple causal models that might reproduce the AI system's behavior. Although this often requires considerable ingenuity, depending on the target domain it might also be possible to refer to prior theoretical and empirical knowledge and apply the tools-to-theories strategy introduced above. Once a causal model has been developed, the variables within this causal model are mathematically "aligned", generally through a search, with the low-level learned variables in the target system. The goal is thus to identify structures within the target network that play the same causal role as the high-level variables in the causal model. Finally, interventions are applied to the causal model and target network to verify that both exhibit the same behavior.

Consider a simple example by Geiger et al. (2021), who aim to explain how a model might perform an arithmetic operation: addition of three integers X, Y, and Z. One very intuitive algorithm for solving this problem is to first add two numbers, X and Y, and then add the sum S_1 to Z, to come up with the final answer S_2 (figure 3). If the network applies this algorithm, it should be possible to localize intermediate encodings that correspond to e.g. S_1 and Z. Given this hypothesized model, the task is to align the internal representations of the target model—the trained neural network—with the nodes of the causal model. If such an "alignment" can be found–and an additional condition of similarity under interventions is met–the causal model can be considered an abstraction of the low-level neural network. While the example above is of course very simple, CAA was also used successfully to explain behavior related to, for example, natural language inference (Geiger et al., 2021).

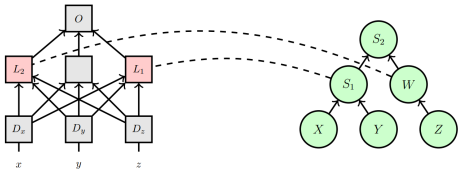

Figure 3: **An example of an alignment between a neural network (left) and a causal model (right).** The locations L_1 and L_2 in the neural network are aligned, for example through a search, to nodes S_1 and W in the causal model.

The main strength of CAA is that it provides a relatively simple model that both reproduces behavior and supports systematic interventions on the original model, through aligning the variables with the simpler, causal model. Moreover, CAA can, in principle, be applied to any network, provided that one has access to its internal variables and parameters. Finally, CAA can be used to an arbitrary degree of precision, depending on the level at which the explanatory causal model is defined.

At the same time, it remains unclear whether CAA scales up to increasingly large and complex AI systems. While it is relatively simple to come up with candidate causal models that explain the addition of three numbers, the scale and number of causal models that could explain more complex behavior like next-word prediction or in-context learning might make this method infeasible in practice. That said, tools-to-theories is known to be a useful guide for constraining the "space" of possible models in cognitive science (Zednik & Jäkel, 2016), and could play a similar role in explainable AI.

Conclusion

To summarize, explainable AI could benefit from taking inspiration from the predominant explanatory strategy in cognitive science. By developing cognitive models to explain the behavior of opaque AI systems, explainable AI can simultaneously facilitate the reproduction of those systems' behaviors and support internal interventions to systematically modify those behaviors. While traditional XAI methods are explanatorily deficient in the sense of falling short of the requirements for trustworthiness—they either do not reproduce behavior or do not provide support for interventions—we argued that two novel methods—mechanistic interpretability and causal abstraction analysis—seem to be adopting a cognitive modeling strategy. Thus, these kinds of strategies, grounded in basic cognitive modeling principles, seem more amenable to ensuring that AI can be trusted.

That said, given that MI and CAA are both very novel methods, some practical challenges remain. For example, it remains an open question whether these methods will scale to large models (e.g. increasingly large LLMs), and whether the provided explanations are sufficient for systematic interventions that can change behavior at a global scale. As such, further methodological and philosophical scrutiny is still necessary to further develop and improve these methods. Even if the connection to cognitive modeling principles is unintentional, these principles provide a revealing lens with which to better understand these methods' explanatory logic, and with which to better evaluate their promise for explainable AI.

References

Bau, D., Zhou, B., Khosla, A., Oliva, A., & Torralba, A. (2017). Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 6541–6549).

Bechtel, W. (2007). *Mental mechanisms: Philosophical perspectives on cognitive neuroscience*. New York: Psychology Press.

Beckers, S., & Halpern, J. Y. (2019). Abstracting causal models. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 33, pp. 2678–2685).

Borowski, J., Zimmermann, R. S., Schepers, J., Geirhos, R., Wallis, T. S., Bethge, M., & Brendel, W. (2020). Exemplary natural images explain cnn activations better than state-of-the-art feature visualization. *arXiv preprint arXiv:2010.12606*.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big data & society*, *3*(1), 2053951715622512.

Busemeyer, J. R., & Diederich, A. (2010). *Cognitive modeling*. Sage.

Craver, C. F. (2007). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. Clarendon Press.

Cummins, R. (1985). *The nature of psychological explanation*. MIT Press.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., ... others (2021). A mathematical framework for transformer circuits. *Transformer Circuits Thread*, *1*.

Geiger, A., Lu, H., Icard, T., & Potts, C. (2021). Causal abstractions of neural networks. *Advances in Neural Infor-*

mation Processing Systems, *34*, 9574–9586.

Gigerenzer, G. (1991). From tools to theories: A heuristic of discovery in cognitive psychology. *Psychological review*, *98*(2), 254.

High-Level Expert Group on AI. (2019). *Ethics guidelines for trustworthy ai* (Report). Brussels: European Commission. Retrieved from https://data.europa.eu/doi/10.2759/346720

Lam, N. (2022). Explanations in ai as claims of tacit knowledge. *Minds and Machines*, *32*(1), 135–158.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Cambridge, MA, USA: MIT Press.

Olah, C., Mordvintsev, A., & Schubert, L. (2017). Feature visualization. *Distill*, *2*(11), e7.

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

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 1135–1144).

Taylor, J. E. T., & Taylor, G. W. (2021). Artificial cognition: How experimental psychology can help generate explainable artificial intelligence. *Psychonomic Bulletin & Review*, *28*(2), 454–475.

Wu, M., Hughes, M., Parbhoo, S., Zazzi, M., Roth, V., & Doshi-Velez, F. (2018). Beyond sparsity: Tree regularization of deep models for interpretability. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 32).

Zednik, C. (2021). Solving the black box problem: A normative framework for explainable artificial intelligence. *Philosophy & technology*, *34*(2), 265–288.

Zednik, C., & Jäkel, F. (2016). Bayesian reverse-engineering considered as a research strategy for cognitive science. *Synthese*, *193*, 3951–3985.