© 2022 The Institute of Mind and Behavior, Inc.
The Journal of Mind and Behavior
Summer 2022, Volume 43, Number 3
Pages 255-268
ISSN 0271–0137

How Can Behavior Be Understood if Its Explanation is Not Comprehended? Does Cognitive Psychology Reach Its Explanatory Limit?

Sam S. Rakover

Haifa University

This essay discusses the attempts of progressive artificial intelligence (AI) models to understand behavior. Because of their sophistication and complexity, it is difficult to understand how these models work and therefore it is difficult to make use of them to understand behavioral phenomena. This is indeed the problem with the present state of cognitive psychology that is founded on the analogy between human behaviors and computer operations: if we do not understand progressive AI models, the most successful and sophisticated programs for predicting behavior, then perhaps this should be interpreted as a warning that cognitive psychology is approaching the limits of its explanatory power. This paper develops and backs up this proposal.

Keywords: explanation, understanding, explainable AI, cognitive psychology

Cognitive psychology, the dominant approach in psychology over the last several decades, is founded on the analogy between computer processing and the processes of sensation, perception, thinking and the like that occur in the mind of an individual. For example, human memory is understood to consist of processes that are parallel to computer processes: input, coding, storage, and information retrieval. However, over the time since the founding of cognitive psychology, information has accumulated that challenges the assumption that the mind operates like a computer. For example, Fodor (2000) raised several arguments that support the claim that the mind's processes do not work like those of a computer. However, no other approach (school) has come to replace the framework of the computer. Even attempts to explain behavior by appealing to research in

I would like to thank two anonymous reviewers and the journal editor, Raymond Russ, for their very useful comments which greatly improve the paper. Correspondence concerning this article should be addressed to Sam S. Rakover, Department of Psychology, Haifa University, Haifa, Israel 31905. Email: rakover@psy.haifa.ac.il

the neurophysiology of the brain are interwoven with the analogy between mind processes and computer information processing.

The problem that arises here is that the most advanced models of AI, "progressive AI models," which successfully predict behavior, are so sophisticated and complex (e.g., there is a huge number of interactions between the enormous quantity of components that compose an AI model) that we cannot understand how they operate or how they generate their outputs; hence they are considered by many computer scientists as "black boxes," i.e., a system that begins with input entering a black box, where the input undergoes some sort of information processing that we do not understand and concludes with the issuing of output that is likewise very difficult to understand.

This situation generates the "explanatory-limit" argument. In brief, the explanatory-limit argument can be presented as follows. Given that (a) currently, the most advanced and successful models for predicting behavior — models based on computational processes — are not understood, and (b) scientific methodology requires that a good theory provides us with accurate explanations and predictions of the behavior under study (e.g., Hempel, 1965, 1966; Keas, 2018; Rakover, 1990, 2018; Salmon, 1990), one may propose two interesting consequences. First, progressive AI models are not good scientific theories, since they do not fulfill the requirement for providing accurate explanations. Second, if one accepts that progressive AI models are currently the best predictive models (theories) that cognitive psychology is capable of producing, then this branch of science runs the risk of approaching its explanatory limits, since we do not understand the behavior that is accurately predicted.

I will now set out a more detailed schema of the explanatory-limit argument. The argument is based on five statements that I take to be correct and which lead to two conclusions:

(a) The goal of cognitive psychology is to explain and understand human behavior.

(b) Cognitive psychology was founded on the analogy of the mind to a computer.

(c) Currently, progressive AI models, which are developed within the framework of cognitive psychology, are the most successful models for predicting behavior. However, these models are not understood and therefore cannot be used for the explanation and understanding of behavior. (Note that these models are incomprehensible to their creators and experts in the studied area, e.g., they are unable to specify the mechanism that produces an output given an input.) It is important to emphasize here that statement (c) refers only to the contemporary situation, and therefore we cannot exclude the possibility that cognitive psychology may develop new and better models in the future capable of providing us with accurate explanations and predictions.

(d) The "explanatory programs" that have been developed to help us understand incomprehensible progressive AI models are limited and have had little success to date.

(e) Of all the theoretical virtues that enable scientists to evaluate an empirical theory, two are essential: explanation and prediction (others include, e.g., consistency and simplicity). Explanation (understanding) is the goal of science, and predictions constitute a major tool for testing theories (e.g., Hempel, 1965, 1966; Keas, 2018; Rakover, 1990, 2018; Salmon, 1990). One may imagine a theory that satisfies such virtues as consistency and simplicity without satisfying the requirements for explanation and prediction (note that Keas, 2018, included both of these in his highest epistemic-weight category: "Evidential virtues").

Conclusion (I): Understanding human behavior by means of these AI models, in the best-case scenario, i.e., using explanatory programs, is limited and partial. Thus, these models do not seem to satisfy the essential requirement for providing an explanation.

Conclusion (II): Currently, it seems that incomprehensible progressive AI models serve to warn us that cognitive psychology is perhaps approaching the limits of its explanatory power. On the one hand, these models exhibit the highest level of predictive accuracy, but on the other hand, they are not comprehended, i.e., we do not understand how they work and therefore we cannot comprehend the studied behavior by using them. However, it should be noted that it is possible that new and better models could be developed within the framework of cognitive psychology in the future to provide us with accurate predictions, explanations, and understanding.

In the present article, I put forward and support what I have called the explanatory-limit argument. In the following section, the boundaries of discussion regarding the paper's principal argument will be outlined.

The Boundaries of Discussion Regarding the Explanatory-Limit Argument

The scope of progressive AI models. The paper will deal only with those incomprehensible progressive AI models (including machine learning and deep neural networks) that attempt to explain human behavior. However, it should be mentioned here that many other similar computer programs, which are not understandable either, are used in other domains such as healthcare, manufacturing, the automobile industry (autonomous vehicles), insurance, banking, and university admissions (e.g., Linardatos et al., 2021; Samek et al., 2017).

My aim is to explore the theoretical implications for cognitive psychology that arise from the finding that progressive AI models are incomprehensible. Therefore, I will not discuss in detail one or two particular AI models, but will concentrate on the ramifications that result from the very fact that progressive AI models are not understood.

"Current state of affairs" approach. As mentioned previously, this paper addresses the current status of progressive AI models, which were designed to explain behavior but are incomprehensible and considered by the literature to

be “black boxes.” I have no idea if, in the future, understandable AI models will be developed and it is difficult, perhaps impossible, to predict whether that will be the case (e.g., Rakover, 2021a). Therefore, the arguments I present here are constrained to the current state of affairs. In an extensive review of the attempts to explain incomprehensible progressive AI models (by using “explanatory programs”), Linardatos et al. (2021) conclude, “Despite its rapid growth, explainable artificial intelligence is still not a mature and well-established field, often suffering a lack of formality and not well agreed-upon definitions” (p. 36). In a similar vein, Gilpin et al. (2019) propose that explanatory programs provide only partial explanations to incomprehensible progressive AI models as they aim at different focal points. Finally, it should be stressed that my focus on the present state of research applies also to other topics relevant to the paper, such as consciousness (e.g., to date, no theory has been developed that explains how consciousness is generated from the brain, for review see Rakover, 2018).

Explanation and understanding. As the above introductory comments make clear, the main arguments of the paper are linked to the concepts of explanation and understanding, and therefore clarifications of these terms are needed (Linardatos et al., 2021, suggest that these concepts are very difficult to define and measure). However, since the literature about these concepts is vast and far beyond the paper’s goal, I will briefly make only two comments that are important and relevant to the present topic of incomprehensible progressive AI models (for reviews of explanation see, for example, Bechtel, 2008; Hempel, 1965, 1966; Rakover, 1990, 2018; Salmon, 1990; Strevens, 2008; Woodward and Ross, 2021).

First, while explanation can be provided by a robot devoid of consciousness, understanding demands human consciousness, i.e., consciousness is a necessary condition to understanding (for arguments supporting this approach see Rakover, 2018, 2021a, 2021b). As an example that illustrates this position, imagine Robbie the robot, the perfect teacher, who can teach classical physics to every student with infinite patience. It turns out that even Robbie’s slowest student eventually understands classical physics and as a result of learning from this perfect teacher, is capable of solving most physics problems with a high score. Should we assume that Robbie the perfect teacher understands his explanations as well as the worst of his students? My answer to that question is no; a robot understands neither the questions nor the answers that it provides its students. (Rakover, 2018, reviews and discusses the claim that even the most sophisticated and complex robot has not developed anything similar to human consciousness.)

Given the importance of consciousness in understanding human behavior, Rakover (2018) reviewed several explanation models presented in the literature and concluded that none of them appropriately treat consciousness as an explanatory concept. Therefore, he developed a new model that takes into consideration the concept of consciousness as an explanation factor (e.g. Rakover, 2011/2012a, 2011/2012b, 2018).

Secondly, the phenomenon under investigation cannot be understood if its explanation is not understood. This is a plain and obvious common sense. Furthermore, lack of understanding of the explanation may lead to an infinite regression. For example, when explanation E1 is offered for not understood phenomenon P, but it turns out that E1 is also not understood, we need another explanation, E2, to explain E1. However, if we do not understand the explanation of the explanation, we need an additional explanation E3... and so on, ad infinitum (here I ignore the possibility that more than one not understood explanation for P is offered). As long as we do not understand the explanation for P, we cannot understand P, and of course, without understanding the explanation, we cannot judge whether the explanation is even partially successful. As can be understood, this idea is at the core of the discussion in this article: the problems related to the lack of understanding of complex computer programs, the progressive AI models.

Organization. The present article discusses the conclusion that progressive AI models are incomprehensible and the attempts to explain them with explanatory programs. This is followed by a discussion of the methodological and philosophical implications of this problem. The final section of the article is a discussion of the consequences of this lack of understanding for cognitive psychology.

Incomprehensible Progressive AI Models and the Attempts to Explain Them by Explanatory Programs

The cases that I will address in this section are related to progressive AI models, including different types of sophisticated and complex software that are used to explain human behavior such as memory, facial recognition and identification, image classification, decision making, language, and categorization (see, for example, Elmahmudi and Ugail, 2019; Gilpin et al., 2019; Kumar, 2021; Linardatos et al., 2021; Samek et al., 2017, 2019; Samek and Muller, 2019; Taylor and Taylor, 2021; Zhou et al., 2019). This type of software, progressive AI models, is based on complicated networks, which contain enormous numbers of components divided between the input layer, the hidden layers (which include a huge number of nodes), and the output layer. However, despite their great success in making predictions, it turns out that understanding them is a big problem. Samek et al. (2017) suggest:

> However, although these models reach impressive prediction accuracies, their nested non-linear structure makes them highly non-transparent, i.e., it is not clear what information in the input data makes them actually arrive at their decisions. Therefore, these models are typically regarded as black boxes. (p. 1)

(Similar comments have been suggested by Linardatos et al., 2021, and Taylor and Taylor, 2021.)

This phenomenon has far-reaching implications, such as mistrust in the validity of the output (decisions, responses, etc.) of the software. Samek and Muller (2019) propose:

> Despite the revolutionary character of this technology, challenges still exist … lack of transparency and explainability, which reduces the trust in and the verifiability of the decisions made by an AI system. (p. 6)

Many of the articles about explainable AI models offer software designed to provide an explanation of progressive AI models, the explanatory programs (for an extensive and thorough review see Linardatos et al., 2021). These kinds of explanatory software offer an explanation, among other things, of the contribution of some groups of nodes in generating the output of the neural network. For example, identifying a cup of coffee or a chicken is based on the detection of groups of nodes that identify the round shape of the cup's opening or the rooster's red crest. In these cases, it can be said that the explanation relies on finding a salient cause for the output (see Samek et al., 2017). Another example is the attempt to identify a face where facial recognition software is trained with partial facial information (as opposed to not training in this way; see Elmahmudi and Ugail, 2019). [It should be noted that the data set with which the network is trained may insert biases into the software. For example, when the training data are based on male responses, the network may learn to prefer a man over a woman in the selection of a candidate for a job, see e.g., Linardatos et al., 2021; Taylor and Taylor, 2021.] Other types of explanatory software use meta-explanations that are based on combining several individual explanations to generate an explanatory pattern, that is, the explanation relies on a schema or generalization as an aid in understanding the output (see Linardatos et al., 2021; Samek and Muller, 2019).

Although some explanatory programs (software) do help to understand progressive AI models to a certain degree, the same disturbing question arises: Do we understand the explanatory programs? This question raises the possibility of an infinite regression of the understanding of the explanation — a point I made earlier.

Samek et al. (2019) write about this matter in the introduction to their book:

> However, many questions remain on whether these explanations are robust, reliable, and sufficiently comprehensive to fully assess the quality of the AI system. (p. v)

Taylor and Taylor (2021) suggest a relatively new approach for solving the problem of incomprehensible progressive AI models. They develop the idea that the research methodology of cognitive psychology can help to discover satisfactory explanations for progressive AI models. This is what they write:

> In this paper, we advance an interdisciplinary approach to XAI (explainable AI) known as Artificial Cognition … drawing heavily on the tradition of experimentation developed within cognitive psychology. This is a call for a new field. (p. 454)

In their paper, they review different types of techniques to explain progressive AI models, discuss their weaknesses, and finally propose the Artificial Cognition approach. I believe that this approach contains a methodological problem. If (a) cognitive psychology's methodology is founded on the research methodology of the sciences (e.g., Taylor and Taylor, 2021, p. 463, propose that it is based on the Popperian falsification approach), and if (b) the research methodology is one that creates incomprehensible progressive AI models, then it is unclear how this methodology will create necessarily successful explanatory software for progressive AI models. In other words, it is not clear how cognitive psychology could help to confer understanding on progressive AI models, because it is founded on the same methodology that generated these incomprehensible models. Nevertheless, it should be stressed that these arguments do not propose that all attempts to explain progressive AI models must be completely unsuccessful. One reason for this, which I would like to emphasize here, is the idea about degrees of understanding (for a development of this idea see Rakover, 2018, 2021b). There are different levels of understanding and one may be satisfied with a low level of explanation (low level of progressive AI model's understanding, e.g., association of certain nodes with particular output). However, if one is interested in a high level of explanation (e.g., a detailed mechanism that generates from specific inputs a specific output), these arguments place a high obstacle on the path to understanding.

The Methodological–Philosophical Implications of Not Understanding Progressive AI Models and Their Explanations

First, I will discuss a case where there is full understanding of an explanation. Second, I will consider the possibility that progressive AI models cannot be explained by using only the mathematical language with which they were created. Finally, I will suggest a possible way to conceive of these incomprehensible softwares as new phenomena.

Full understanding of the explanation. Imagine a seventeenth-century scholar of human behavior who is deeply impressed by Newton's mechanistic approach to solving physics problems. Suppose he has adopted a theoretical approach that a perfectly mechanistic explanation of human behavior is possible. As a way of supporting and demonstrating his behavioral–mechanistic theory, he builds Isaac the robot, who can perfectly imitate relatively simple human behaviors: he can pour a cup of tea and sign his name on a piece of paper. The mechanism that performs these behaviors is made of springs, metal shafts, and wires, gears, weights, etc. The explanation of this is straightforward. One first needs to wind up the spring in Isaac's back. Then, one must pull the appropriate handle for signing its name or pouring the cup of tea, activating the mechanism. It is possible to explain the operation of the mechanism using a schematic diagram that precisely describes every movement of every part of the robot that together cause it to sign its name or pour the cup of tea. This precise and detailed description of the signature mechanism is the complete explanation of the behavior of the robot that can be understood by anyone. However, can Isaac the robot understand its own actions? It obviously understands nothing, even though a human could understand.

This example has two important implications. First, it indicates that, like Isaac the robot, progressive AI models do not understand what they are doing, because all these softwares lack consciousness (Rakover, 2018, argued that no computer has developed consciousness). Second, while human beings can understand how Isaac works, they cannot understand the very complex actions performed by progressive AI models or their complex explanatory programs. At this point, we must ask ourselves: How is this possible? Was this software not written by programmers who must have understood what they were creating? How then is it possible that no one understands what these programs are doing? A possible answer is this: one may conceive of these programs as broad frameworks within which different enormous complex series of events that require explanation take place. The principles by which the progressive AI models were designed are insufficient to explain these series of events. This idea can be explicated by the analogy to chess.

Nearly everyone knows the rules of chess and nearly everyone has played this beautiful game at one time or another. However, although these rules are what distinguish chess from other board games like checkers and backgammon, it is impossible to explain why Mikhail Botvinnik was one of the greatest chess players just by appealing to the game's rules. To understand how Botvinnik was a dominant player we need to take a number of factors into account that are not directly related to the rules of the game, like his mastery of strategy and tactics (openings and end game), his ability to grasp a game situation in an instant, his ability to think ahead to future moves, his nerves of steel and his understanding of his opponents' style of play. Programming progressive AI models is analogous to fixing the rules of the game, within which the program learns to play and to perform actions, that is, to achieve certain goals like facial recognition, decision making, and the categorization of objects. In other words, I suggest that the series of equations that programmers use in order to create progressive AI models are no more than the rules that set up the framework within which a program will develop in such a complex way that it will be very difficult to understand. The fact that there is no clear answer to the question as to how exactly the program learns and develops testifies to the fact that a progressive AI model is a "black box." It is for this reason that we need explanatory programs to explicate these opaque models. Let us call this the "new emergent phenomenon."

A progressive AI model as a new emergent phenomenon and different levels of understanding. The fact that progressive AI models are not understood inspires the production of explanatory programs (software) as well as studies that use experiments to decipher what they are doing (see, for example, Elmahmudi and Ugail, 2019; Samek et al., 2019; Taylor and Taylor, 2021; Zerilli, 2022). These models can be conceived of as new phenomena that need to be explained, i.e., the progressive AI models themselves have become objects of interest. How are we to relate to the fact that these models are a new phenomenon that needs to be understood by explanatory programs?

I will discuss two aspects of this question and end with the "explanatory-program" argument. The first point refers to degrees of understanding. It is possible to say that although an explanatory program (software) of progressive AI models does not provide full understanding, it does provide partial, imperfect explanations. For example, Samek et al. (2017) compare two types of explanatory software, sensitivity analysis (SA), and layer-wise relevance propagation (LRP), and find that the explanations provided by LRP are better than those provided by SA. Likewise, Linardatos et al. (2021) analyzed several explanatory programs and came to the conclusion that two of them "…are, by far, the most comprehensive and dominant across the literature methods for visualizing feature interactions and feature importance…" (p. 35). In view of this, it may be proposed that these AI models provide a limited level of understanding that is anchored in (a) an explanatory program, and (b) the way the progressive AI models were programmed. As an example for (b), one can achieve some understanding of how certain machine learning models were developed and trained by appealing to the special algorithm used by the designers called "backpropagation": in short, this algorithm uses an error made by the software (the gap between the output value and the behavioral value) to change the weightings (the strengths of the connections between the nodes that constitute the model) so that this gap will gradually shrink and the power of the neural network to predict the behavior under investigation will gradually increase.

The second point refers to the difference between a natural phenomenon and an incomprehensible progressive AI model as new phenomena. An incomprehensible progressive AI model is an explanatory model of behavior and as such, it may be incorrect, while a natural phenomenon is neutral in this respect. If the incomprehensible progressive AI model is incorrect (and an incorrect theory may generate correct predictions), then an explanatory program will cheat a human user when it does not point out that the incomprehensible progressive AI model is wrong (e.g., it is biased). Moreover, while it is relatively easy to empirically test a scientific theory of a natural phenomenon (e.g., by falsification), it is hard and very complicated to test an explanatory program when the progressive AI model (as a new phenomenon) is incomprehensible.

In view of all this, the following abstract argument, the explanatory-program argument, may be developed. Let us assume that T^* signifies an incomprehensible progressive AI model, a theory of behavior B, and T signifies an understandable explanatory program, a theory that attempts to explain how T* works. In what sense can this attempt be understood? T cannot be a reducing theory of T*, since T* is not comprehended. Furthermore, T cannot be a justification for T*, since justification is not a scientific explanation (however, see Zerilli, 2022, who suggests a theoretical justification for an incomprehensible machine learning). Given this situation, how can one scrutinize T? One important way is to test the following prediction derived from T: T* in condition C will predict a particular behavior B. If this prediction is refuted, then T cannot be conceived of as a satisfactory theory. However, if T is successful, then one may give up T*, since T does the job of T* and moreover *T is understandable*. Thus, T* brings us near the limit of scientific goal, which is prediction without understanding. On the one hand, T* predicts accurately behavior B, but since we do not understand it we do not understand behavior B. On the other hand, if we develop T, which predicts everything that T* predicts and also explains these predictions, then T* becomes needless. Thus, it seems that there are no other options but to develop a theory that fulfills both crucial demands of scientific theory: prediction and comprehension. This conclusion is in line with a previous discussion of theoretical merits, in which Keas (2018) included prediction and explanation in his highest epistemic-weight category: “Evidential virtues.”

Discussion

This discussion concentrates on two important issues. First, the connection between the analogy of cognitive psychology and progressive AI models, and second, a proposed answer to the question of why these models are so hard to understand, particularly in the realm of human behavior.

Analogy. Analogies are an important tool for the explanation of behavior. Let us explore the following schema that characterizes cognitive psychology. If we conceive of human behavior in general in the following way:

Response (Y) = f [Unknown Mechanism, stimulus (X)]

and if we find some device, like a computer or Isaac the robot that behaves in the following manner: Response (Y*) = f [Known Mechanism, stimulus (X*)], where response (Y) is very similar to response (Y*) [e.g., pouring a cup of tea], and where stimulus (X) is very similar to stimulus (X*) [the situation in which tea is poured], then we will tend to reach the conclusion that the unknown mechanism in the human is very similar to the known mechanism in the computer or in the robot.

Two comments should be made about this analogy. First, the fact that two things, each made out of many different components, exhibit significant resemblances with regard to some specific set of components does not ensure that significant resemblances will be found in other components. As mentioned above, there are important (functional) similarities between the behavior of a computer and a person: between the input and the stimulus and between the output and the response; additionally, there are similarities between several subsystems in a computer and certain subsystems in the human brain. Despite these similarities, it is easy to point out the vast differences between the functioning of a computer and human cognitive functioning. For example, in many areas, a computer's computational power is greater than that of a human by several orders of magnitude while a computer has not yet generated consciousness like a human (for other differences between humans and machines, see Borowski et al., 2021; Fodor, 2000).

The similarity of the actions of pouring tea or signing a name between Isaac the robot and a human person does not necessarily mean that the mechanism responsible for the robot's actions is the same as the mechanism responsible for the person's actions. In this case, it is entirely clear that they are completely different mechanisms. The logical reason why the analogy does not necessarily assure a correct explanation is anchored in the following fact: every data set can, in principle, be derived from an infinite number of different functions (i.e., theories). In this case, the data connected to the state of affairs for the signing and the response of signing or the state of affairs for the pouring of the tea and the response of pouring the tea involve two different mechanisms, one entirely mechanical and the other physiological, cognitive, and mental.

Secondly, the analogy is especially tempting when not understood behavior A is compared to understood behavior B. In that case, we tend to apply the explanatory-mechanism of B to the not understood behavior A. However, when explanatory-mechanism B is itself not understood, the use of activity B as an analogical explanation of A becomes problematic. As a matter of fact, that is the present state of cognitive psychology: progressive AI models are incomprehensible.

Since we currently fail to understand progressive AI models, which are the most successful models for predicting behavior, the explanatory-limit argument set out and supported here is particularly salient: if at present we do not understand the explanations these models generate, then we cannot understand the behavior that they were created to elucidate in the first place. Therefore, it may be suggested that (a) progressive AI models, which were generated for understanding human behavior, have thus far failed in their mission to offer the required explanations, and (b) since these progressive AI models are incomprehensible (even if they are currently considered to be the best models of predicting human behavior), one may question whether the foundation of cognitive psychology, which is based on the computer–mind analogy, is solid (for arguments against this analogy, see Fodor, 2000).

In other words, it may be proposed that progressive AI models are like the incomprehensible explanatory-mechanism in the analogy between behaviors A and B, which therefore indicates that cognitive psychology is perhaps approaching its limits of explanatory power: although there are many similarities between human behavior and a computer's functions (the progressive models predict human behavior accurately), we encounter an obstacle in applying the explanatory mechanism of these most advanced computer models to human behavior — their explanatory mechanism is not understood. And if we currently fail to understand the explanatory mechanism of these AI models, one may issue a warning: perhaps this analogy, on which cognitive psychology is based, is weak at the explanation level. Once again, one should be cautious here, since it is possible that cognitive psychology may create good models in the future capable of providing satisfactory explanations and predictions.

The above discussion raises the problem about the methodological importance of prediction vs. explanation in cognitive psychology, because of the following consideration. Progressive AI models are one of cognitive psychology's highest-level theories, yet they are not capable of providing us with a high-level explanation of behavior. It appears that the greatest strength of these models lies in their predictive power: their outputs match empirical observations very well. However, it is precisely this strength that raises a difficult methodological problem: it is here that the gap between prediction and explanation can be seen. The methodological emphasis moves from explanation to prediction (i.e., the accuracy of predicting outcomes). The question that arises here is whether such a methodological move is beneficial to psychology. My answer is that it is not. At the extreme, this shift from an emphasis on explanation to prediction results in the acceptance of any incomprehensible theory (e.g., the progressive AI model) as long as it successfully predicts the observed results. This prediction approach could lead to a dramatic decline in the quality of scientific research; without scientific understanding, we will not be able to construct empirical tests for theories and models. There will be no theoretical basis for making a specific prediction given a certain condition. We will have a difficult time distinguishing between many possible theories that fit the same observed results.[1] In fact, without understanding, the well-predicting theories will become ad hoc, since it will be impossible to test them empirically. This argument is founded on the common-logic notion that a true explanation produces a correct successful prediction (i.e., that under the relevant conditions, the correct theoretical explanation will generate a correct prediction). However, successful prediction is only a necessary condition of a correct explanation. A correct explanation cannot produce a false prediction, but an incorrect theory can produce a correct prediction.

[1] Keas (2018) puts forward twelve theoretical virtues which can be used to evaluate and decide between theories. However, it should be emphasized that Keas' goal, the systematization of virtues, is not the topic of the present paper. The dispute here is about the question of which of the two theoretical virtues is more important for cognitive psychology: explanation or prediction?

The present approach contradicts the one that emphasizes the crucial importance of predictions over explanation in psychology. Yarkoni and Westfall (2017) present several examples based on machine learning that support the "prediction priority" approach and conclude "Our argument has been that psychologists stand to gain a lot by relaxing their emphasis on identifying the causal mechanisms governing behavior and focusing to a greater extent on predictive accuracy" (p. 1118). Methodologically, I believe that the prediction-priority approach is overstated, since, as mentioned above, without explanation and understanding scientific progress will be stopped and ruined. Yet, without predictions (to be compared with observations) science cannot progress either, since prediction is an essential component in theory testing. So, in accordance with the above discussion, I would propose that explanation (understanding) and prediction (observation) are both necessary conditions for scientific progress.

Progressive AI models and human behavior. The question of why progressive AI models are not understood has been answered above by emphasizing two factors: the tremendous complexity of these models, and our perception of them as a new emergent phenomenon that cannot be apprehended by appealing only to the mathematical tools which generated these models in the first place. However, since the present paper deals with progressive AI models that attempt to comprehend human behavior, let me propose an additional crucial factor that may explain our difficulty in understanding these models. The crucial factor is consciousness. The main argument for this can be stated simply: while any attempt to understand human behavior involves consciousness, this crucial factor is missing from progressive AI models. Justifications of this claim have been developed by Rakover (2011/2012a, 2011/2012b, 2018, 2021a, 2021b). Very briefly, the following can be said. In setting out the analogy between a computer and brain activity and function, it has become apparent that while computer function is mechanistic, i.e., its theoretical framework is based on concepts used in the natural sciences — concepts that are not influenced at all by mental processes such as desire and belief — human functioning is mentalistic and includes conscious experience like desire and belief which are intertwined with behavior. As long as we do not have a theory that explains how consciousness is grounded in neurophysiology, it is difficult to see how progressive AI models, which are founded on the mechanistic frame of reference of computers, can offer us a full explanation of the conscious behavior of humans. In other words, it is hard to see how the explanatory mechanism of these models mirrors the explanatory role that consciousness plays in the explanation and understanding of human behavior.

References

Bechtel, W. (2008). *Mental mechanisms: Philosophical perspectives on cognitive neuroscience*. LEA.

Borowski, J., Stosio, K., Brendel, W., Wallis, T. S.A., and Bethge, M. (2021). Five points to check when comparing visual perception in humans and machines. *Journal of Vision*, *21*, 1–23.

Elmahmudi, A., and Ugail, H. (2019). Progressive face recognition using imperfect facial data. *Future Generation Computer System*, *99*, 231–225.

Fodor, J. (2000). *The mind doesn't work that way: The scope and limits of* computational psychology. MIT Press.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A. Spector, M., and Kagal, L. (2019). Explaining explanations: An overview of interpretability of machine learning. *ArXiv:1806.00069v3 [cs.AI]*.

Hempel, C. G. (1965). *Aspects of scientific explanation and other essays in the philosophy of science*. The Free Press.

Hempel, C. G. (1966). *Philosophy of natural science*. Prentice–Hall.

Keas, M. N. (2018). Systematizing the theoretical virtues. *Synthese*, *195*, 2761–2793.

Kumar, A. A. (2021). Semantic memory: A review of methods, models, and current challenges. *Psychonomic Bulletin & Review*, *28*, 40–80.

Linardatos, P., Papastefanopoulos, V, and Kostsiantis, S. (2021). Explainable AI: A review of machine learning interpretability methods. *Entropy*, *23*, 1–45.

Rakover, S. S. (1990). *Metapsychology: Missing links in behavior, mind and science.* Paragon/Solomon.

Rakover, S. S. (2011/2012a). A plea for methodological dualism and multi-explanation framework in psychology. *Behavior and Philosophy*, *39/40*, 17–43.

Rakover, S. S. (2011/2012b). Methodological dualism and multi-explanation framework: Replies to criticisms and further developments. *Behavior and Philosophy*, *39/40*, 107–125.

Rakover, S. S. (2018). *How to explain behavior: A critical review and new approach.* Lexington Books.

Rakover, S. S. (2021a). *Understanding human behavior: The innate and acquired meaning of life.* Lexington Books.

Rakover S. S. (2021b). Two factor theory of understanding (TFTU): Consciousness and procedures. *Journal of Mind and Behavior*, *42*, 347–370.

Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world.* Princeton University Press.

Salmon, W. C. (1990). *Four decades of scientific explanation*. University of Minnesota Press.

Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K. and Muller, K-R. (Eds.). (2019). *Explainable AI: Interpreting, explaining and visualizing deep learning.* Springer Nature Switzerland AG.

Samek, W., and Muller. K-R. (2019). Toward explainable artificial intelligence. In W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K-R. Muller (Eds.), *Explainable AI: Interpreting, explaining and visualizing deep learning* (pp.1–22). Springer Nature Switzerland AG.

Samek, W., Wiegand, T., and Muller, K-R. (2017). Explainable artificial understanding, visualizing and interpreting learning model. *arXiv: 1708.08296v1.*

Strevens, M. (2008). *Depth: An account of scientific explanation*. Harvard University Press.

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

Woodward, J., and Ross. L. (2021). Scientific explanation. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy.* <https://plato.stanford.edu/archives/sum2021/entries/scientific -explanation/>

Yarkoni, T., and Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*, 1100–1122.

Zerilli, J. (2022). Explaining machine learning decisions. *Philosophy of Science,* 89, 1–19.

Zhou, B., Bau. D., Oliva, A., and Torralba, A. (2019). Comparing the interpretability of deep networks via network dissection. In W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K-R. Muller (Eds.), *Explainable AI: Interpreting, explaining and visualizing deep learning* (pp. 243–252). Springer Nature Switzerland AG.