How can Behavior be Understood if its Explanation is not Comprehended? A Methodological Critique of Cognitive Psychology

 Sam S. Rakover

Department of Psychology, Haifa University, Haifa, Israel 31905

Running head: Critique of Cognitive Psychology

Rakover phone number: 972 4 8240924

Email: rakover@psy.haifa.ac.il

Correspondence should be addressed to Sam S. Rakover, Department of Psychology, Haifa University, Haifa, Israel 31905.

Email: rakover@psy.haifa.ac.il

# Abstract

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 the progressive AI models, the most successful and sophisticated programs at predicting behavior, **this might be taken to suggest that cognitive psychology has reached the limit of its explanatory power. This paper develops and backs up this proposal.**

*Keywords*: explanation, understanding, explainable AI, cognitive psychology

**How can Behavior be Understood if its Explanation is not Comprehended? A Methodological Critique of Cognitive Psychology**

# The Critique Argument

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. It seems that the belief that the mind is operating like a computer has been weakened from the time cognitive psychology was established. For example, Fodor (2000) raised several arguments that support the claim that mind’s processes do not work like a computer’s processes. However, yet I don’t see any new approach (school), which replaces the framework of the computer. Even the attempts to explain behavior by appeal to research in the neurophysiology of the brain is interwoven with the analogy between mind processes and computer’s information processing.

The problem that arises here is that the most advanced models of AI for understanding human behavior, “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 the model) that we cannot understand how they operate or how they generate their outputs; hence they are considered by many computer scientists black boxes, i.e., systems that begin with input entering a black box, where it undergoes some sort of information processing that we do not understand and concludes with the issuing of output that is very difficult to understand.

This situation generates the “critique argument.” In brief, the critique argument can be stated as follows: The fact that we cannot understand progressive AI models immediately raises the question of the title of the article: How do you understand behavior if you do not understand its explanation? **If the most advanced and successful models for predicting behavior – models based on computing processes – are not understood, then one might suggest the following. Progressive AI models hint that cognitive psychology has reached its “explanatory limit,” since we do not understand the model that accurately predicts the studied behavior and therefore do not comprehend that behavior. This casts doubt on whether cognitive psychology can provide a complete explanation of human behavior and thus achieve its goal of understanding it.**

I will now set out a more detailed schematic of the critique argument. The argument is based on the five statements that I take to be true which lead to two conclusions:

1. The goal of cognitive psychology is to explain and understand human behavior.
2. Cognitive psychology is founded on the analogy of the mind to a computer.
3. The most important computer models, which are successful in predicting behavior, are developed within the framework of cognitive psychology for understanding human behavior are progressive AI models. **(It is important to note that cognitive psychology includes a large number of models, whose predictive ability is less than that of progressive AI models.)**
4. Progressive AI models, which successfully predict behavior, are not understood and therefore they cannot be used for the explanation and understanding of human behavior. **(These models are incomprehensible to their creators and experts in the studied area.)**
5. The “explanatory programs” that have been developed to help us understand incomprehensible progressive AI models are only a limited and partial success.
6. Therefore, understanding human behavior by means of these AI models, in the best-case scenario, i.e., using explanatory programs, is only limited and partial (sometimes that understanding can even be misguided – see below).
7. **Therefore, one might propose the following. Perhaps progressive AI models are approaching the limit of cognitive psychology’s 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 do not comprehend the studied behavior.**

This argument is explicated, elaborated, and supported in the present paper. In the following section, major topics and concepts used in this paper are clarified and the organization of the article is described.

# Some Clarifications of the Paper’s Topics, Concepts and its Organization

*The scope of progressive AI models*: While the paper deals with those incomprehensible progressive AI models (including machine learning and deep neural nets) that attempt to explain human behavior, many similar computer programs 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). Although many of these programs are also not understandable and are therefore sometimes mistrusted, generating a demand for explanation, these other similar computer programs are beyond the scope of this article.

My aim is to explore for cognitive psychology the implications 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 concentrate on the ramifications that result from the very fact that progressive AI models are not understood.

*“Present state” approach*: This paper addresses the current state of progressive AI models, which were designed to explain behavior but are incomprehensible and considered by the literature ‘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, 2021). Therefore, the arguments I present here are limited 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 only 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*: Since the main arguments of the paper are linked to the concepts of explanation and understanding, 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 make briefly only the following three important and relevant comments (for reviews of explanation see for example, Bechtel, 2008; Hempel, 1965, 1966; Rakover, 1990, 2018; Salmon, 1990; Strevens, 2008; Woodward & 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.)

Second, Rakover (2018) reviewed several explanation models described in the literature and concluded that none of them handle appropriately consciousness as an explanatory concept. He developed a new model that take in to consideration the concept of consciousness as an explanation factor (e.g. Rakover, 2011/2012a, 2011/2012b, 2018).

Third, the phenomenon under investigation cannot be understood if its explanation is not understood. This is a plain and obvious common sense: without understanding the explanation, one cannot understand the phenomenon under investigation. 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 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 & Ugail, 2019; Gilpin et al, 2019; Kumar, A., 2021; Linardatos et al, 2021; Samek, Montavon, et al, 2019; Samek, & Muller, 2019; Samek, Wiegand, et al., 2017; Taylor & Taylor, 2021; Zhou, Bau, 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, Wiegand, et al. (2017) write:

“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)

Similarly, Taylor & Taylor (2021) write:

“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.” (p. 454)

Likewise, Linardatos et al. (2021) write:

“However, this improved predictive accuracy has often been achieved through increased model complexity. … As a consequence, the rationale behind their decisions becomes quite hard to understand and, therefore, their predictions hard to interpret.” (p. 1)

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

“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, Wiegand, 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). In this case, one may also suggest that a salient cause was found, the part of the face, for the facial recognition (see Elmahmudi & 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 is 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 & 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, & 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, Montavon, 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 & 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 the following possible methodological problem. If (a) cognitive psychology's methodology is founded on the research methodology of the sciences (e.g., Taylor & Taylor, p. 463, propose that it is based on the Popperian falsification approach), and if (b) the sciences’ research methodology is the one that creates incomprehensible progressive AI models, then it is unclear how this very 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). 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). However, if one is interested in a high level of explanation, these arguments place a high obstacle on the path to understanding.

# The Methodological-Philosophical Implications of not Understanding Progressive AI Models and their Explanations

I will first discuss a case where there is full understanding of an explanation, and then I will discuss a possible way to handle incomprehensible software as a new phenomenon.

# 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 Robert 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 the robot’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 Robert 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 Robert the robot, not only do progressive AI models not understand what they are doing, even their explanatory programs do not understand what they are explaining, because all of these types of software lack consciousness (Rakover, 2018, argued that no computer has developed consciousness). Second, while human beings can understand how Robert 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? Wasn’t this software 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? There are two parts to the answer. First, the lack of understanding is due to the vast and profound complexity of the progressive AI models. Second, one may conceive of these programs as broad frameworks within which 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 & Ugail, 2019; Taylor & Taylor, 2021; Samek, Montavon, et al. 2019). 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 that we seek to understand. 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. First, there is a difference between a natural phenomenon and an incomprehensible progressive AI Model as new phenomena. An incomprehensible progressive AI model is a model of reality 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 a good explanatory program will deceive 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.

 The second 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, Wiegand, 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 by SA. Furthermore, Linardatos et al. (2021) suggest that “the LIME and SHAP methods are, by far, the most comprehensive and dominant across the literature methods for visualizing feature interactions and feature importance…” (p. 35). It is clear from these examples that many different types of explanatory software do not provide complete explanations to not-understood progressive AI models but only partial explanations. This raises the following point: despite the difference, mentioned above, between a scientific theory and an explanatory program with regard to its capacity to be empirically tested, one may propose that an explanatory program is similar to a scientific theory in that the latter also provides partial explanations. There are several factors responsible for this. Here I will mention two reasons that scientific theories provide only partial explanations:

1. *Confirmation and Falsification*: Every empirical theory is provisional and is considered confirmed until it is falsified (Popper, 1972; Rakover, 2018). For example, Newtonian mechanics was considered a correct theory until it became clear that some of its basic assumptions do not apply when a body’s velocity approaches the speed of light.
2. *The Scope of the Explanation*: Every empirical theory is limited, either explicitly or implicitly, by a theoretical boundary. The explanations apply to all relevant phenomena within that scope. For example, Newtonian mechanics applies to all bodies moving at earthly velocities (that do not approach the speed of light). Another example: Is Robert the robot, as described above, a good candidate for explaining human behavior? Intuitively, the answer is no, and anyone can think of a number of significant differences that relate to the scope of Robert’s actions compared to the scope of human action: Robert the robot moves by the action of springs, rods, and gears while a human being moves by utilizing bones, muscles, and nerves; a human being, in contrast to Robert, adjusts his actions to his environment due to his possession of consciousness. A final example: Rakover & Cahlon (1989, 2001) developed a mathematical model, “the Catch model,” for the reconstruction of a target face from a witness’s memory. The model’s assumptions preclude the interfering effects on memory of viewing other faces. Under these assumptions, it has been proven mathematically that a target face can be reconstructed from a witness’s memory. However, since exposure of the witness to additional faces is part of the model’s procedure for facial reconstruction, strong interference with the witness’s memory actually occurred. In other words, if indeed a witness’s memory of faces were not negatively affected by additional facial information, the Catch model would be successful at reconstructing a target face from an individual’s memory.

On the reasonable assumption that every empirical theory provides a certain level of understanding of the phenomenon being investigated, one may offer the following suggestion. On the one hand, one may conceive of progressive AI models as so complex that they are not understandable; they are black boxes. On the other hand, one may propose 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 appeal 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.

# Discussion

**This discussion will concentrate 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 Robert 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 appropriate human is very similar to the known mechanism in the computer or the robot.

Two comments should be made about this analogy. First comment: 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 Robert 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.

Second comment: 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: if we do not understand progressive AI models, which are the most successful models at predicting behavior, the critique argument set out and supported here becomes salient: **if we do not understand the explanation, then we do not understand the behavior for which that explanation was given in the first place.**

**Given this, it may be proposed that progressive AI models are like the incomprehensible explanatory-mechanism analogy (between behaviors A and B), and therefore indicate that cognitive psychology has reached its “explanatory limit”: 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, because this mechanism is not understood. And if we do not understand the explanatory mechanism, which can be viewed as the basis of cognitive psychology’s analogy, it may be proposed that this analogy has reached the limit of its explanatory power.**

**The above argument 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., 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. 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.

The present approach contradicts the one that emphasizes the crucial importance of predictions over explanation in psychology (see Yarkoni & Westfall, 2017). In their article, Yarkoni & Westfall 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, I would propose that explanation (understanding) and prediction (observation) are both necessary conditions for scientific progress.

***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 their perception as a new emergent phenomenon. 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. I believe that this 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.** **Rakover (*ibid*.) has developed within the accepted scientific methodology a special explanatory framework that combines both mechanistic and mentalistic approaches.**

# References

# Bechtel, w. (2008). *Mental mechanisms: Philosophical perspectives on cognitive neuroscience*. New York: LEA.

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

Elmahmudi, A. & 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*.Cambridge, MA: MIT Press.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A. Spector, M. & 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.

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

Linardatos, P., Papastefanopoulos, V, & Kostsiantis, S. (2021). Explainable AI: A review of Machine learning interpretability methods. *Entropy*, 23, 18. P. 1-45. http://dx.doi.org/10.3390/e23010018

Popper, K. R. (1972). *Objective Knowledge: An Evolutionary Approach.* Oxford University Press.

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 Sam S. (2021b). Two Factor Theory of Understanding

(TFTU): Consciousness and procedures. *The Journal of Mind and Behavior*, 42, 347-

370.

Rakover, S. S., & Cahlon, B. (1989). To catch a thief with a recognition model: The model and some empirical results. *Cognitive Psychology*, 21, 423-468.

Rakover, S. S., & Cahlon, B. (2001). *Face recognition: Cognitive and computational processes*. John Benjamins.

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. & Muller. K-R. (2019). Toward explainable artificial intelligence. In Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K. & Muller. K-R (Eds.)

*Explainable AI: Interpreting, explaining and visualizing deep learning,* (pp.-22).Springer Nature Switzerland AG.

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

Springer Nature Switzerland AG.

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

Strevens, M. (2008). *Depth: An Account of Scientific Explanation*. Cambridge Mass.: Harvard University press.

Taylor, J. E. T. & Taylor, G. W. (2021). Artificial cognition: How experimental

psychology can help generate explainable artificial intelligence. *Psychonomic Bulletin and Review*, 28, 454-475.

Woodward, J, & Ross. L. (2021). Scientific explanation. In E. N. Zalta (ed.), The Stanford Encyclopedia of Philosophy. URL = <https://plato.stanford.edu/archives/sum2021/entries/scientific-explanation/>.

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

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