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: Incomprehensible AI models

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 progressive Artificial Intelligence (AI) computer models that attempt to understand human behavior. However, because of their sophistication and complexity, it is difficult to understand how these models work and in the absence of that understanding, it is difficult to make use of them to understand the 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 that predict behavior, it is not clear whether cognitive psychology will achieve its goal: to explain human behavior. The article discusses this question while briefly describing the basic principles of explanation models, the problem of lack of understanding of progressive AI models and the attempts to explain them, the methodological-philosophical implications of this problem, and, finally, the repercussions of this lack of understanding for cognitive psychology.

*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 (or even an animal). For example, human memory is understood to consist of processes that are parallel to computer processes: input, coding, storage, and information retrieval.

**The problem that arises here is that the most advanced models of AI for understanding human behavior, “progressive AI models”, are so sophisticated and complex (e.g., there are huge amount of interactions among enormous quantity of components that compose the model) that we cannot understand how they operate, how they generate the outputs; they are actually considered 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 conclude with the issuing of output that is very difficult to understand.**

**This situation creates the critique argument. I will first describe it briefly and then in more details.** The finding that progressive AI models do not provide with understanding of behavior 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, models based on computing processes, are not understood, then how can cognitive psychology achieve its purpose of explaining behavior? This is a serious obstacle to the success of the cognitive psychological approach.

**Considering this, I will now describe the critique argument in more detail. Given that (a) Cognitive psychology is founded on the computer analogy,**

**(b) The goal of cognitive psychology is to explain and understand human behavior,**

**(c) Progressive AI models for understanding human behavior are considered as most important models developed within the framework of cognitive psychology,**

**(d) Progressive AI models are not understood and therefore they cannot be used for explanation and understanding human behavior,**

**(e) Since “explanatory programs” that are designed for understanding the incomprehensible progressive AI models have only a limited and partial success, understanding of human behavior by these AI models can only be limited and partial in the best case scenario (sometimes understanding may be even misguided – see below),**

**Given these, one may cast doubt if cognitive psychology will achieve its purpose of explaining and understanding human behavior.**

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

**Some clarifications of the paper’s topics and its organization**

***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 deal with other domains such as healthcare, manufacturing, automobile industry (autonomous vehicles), insurance, banking, and admission to university (e.g. Linardatos et al, 2021; Samek et al, 2017). Although in many cases the latter programs are not understandable and therefore they are mistrusted and a demand for their explanation is created, I will not refer to them in particular. The reason is that my aim is to explore the ramifications of incomprehensible progressive AI models to cognitive psychology.**

***“Present-state” approach*: The paper handles the current state of progressive AI models, which were designed to explain behavior but are incomprehensible and considered ‘black boxes’. I have no idea if in the future understandable AI models will be developed. In my view this prediction is hard if not impossible to make (e.g., Rakover, in press). So, the arguments I present here are relevant to the current state of affairs. In an extensive review of the attempts to explain inapprehensible 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 inapprehensible progressive AI models as they aim at different focal points. Finally, it should be stressed that the present-state approach includes other topics relevant to the paper, such as consciousness (e.g., to date, no theory has been developed that explain how consciousness is generated from the brain, see Rakover, 2018).**

***Explanation and understanding*: Since the main arguments of the paper are anchored to the terms of explanation and understanding, an attempt will be made to clarify them (Linardatos et al, 2021, suggest that these concepts are very difficult to define and measure). The proposed clarification is based on the following important notions. First, explanation and understanding are ‘not all or nothing’ concepts. In my view they come in degrees. Second, 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, in press). Third, since one may conceive of an incomprehensible progressive AI model as a black box, i.e., as a new phenomenon to be explained, a short review of scientific explanation may seem contributory: it may help to grasp in what sense the explanatory programs illuminate inapprehensible progressive AI models.**

 ***Organization*:This article includes a brief summary of the basic assumptions of scientific explanation models; the problem of incomprehensibly progressive AI models and the attempts to explain them by explanatory programs; a discussion of the methodological and philosophical implications of this problem; and a discussion of the consequences of this lack of understanding for cognitive psychology.**

# The Basis of an Explanation Model

Modern science is characterized by the attempt to achieve the important goal of explaining a phenomena rationally. The nature and importance of this project have gradually become clearer since the days of Copernicus and Galileo. Explanation involves fulfilling certain methodological requirements that are widespread in the sciences (including the social sciences). For the purposes of this article, I list the following methodological requirements (see Rakover, 2018):

1. For the explanation of a phenomenon, one must use an explanation model: a special procedure that is appropriate to the type of phenomenon investigated.

This requirement is based on the fact that the phenomenon that has been observed meets the appropriate methodological requirements (the phenomenon must meet the requirements of objectivity, repeatability, and publicity, see Rakover, 1990). The observation is represented in a particular language (for example, mathematical representation) and is explained by a hypothesis, a model, or a relevant theory. The theory cannot explain the phenomenon in question by itself. For explanation to be possible, the proper explanation procedure must be applied to the theory. For example, to explain the distance a body will travel in a free fall, one uses Hempel’s (1965, 1966) deductive-nomological (D-N) model: From the two premises, the law of free-falling bodies (Galileo’s law) and a specific given time, one logically (mathematically) deduces a particular result. In other words, one predicts the distance that the body will travel in a given time. If it is empirically determined that the body has indeed traveled the predicted distance, then the phenomenon is explained by this law which is based on gravity (there are, of course, other procedures, explanation models that I will not discuss here, see Rakover, 2018).

**The entire range of explanation procedures that appear in the literature can intuitively be categorized into four main types of explanation, which give answers to *why* and *how* questions (see discussion in Gilpin et al, 2019; Linardatos et al, 2021; Rakover, 2018; Salmon, 1984, 1990. Here I avoid discussing certain ethical and legal issues that are relevant to inapprehensible progressive AI models and their explanations):**

## *A particular instance*. The explanation procedure shows that the phenomenon under discussion is a particular instance of a law or general theory (regardless of whether the theory is deterministic or probabilistic). For example, an iron ball falling 4.9 meters in the first second of free fall is a particular instance of Galileo’s law (which can be explained by the Newtonian theory).

## *Causation.* The explanation procedure shows that the phenomenon can be understood by inserting it into a causal web. For example, Mr. Hashimoto contracted leukemia due to his exposure to radioactive radiation that resulted from the destruction of the Fukushima nuclear plant by a tsunami.

## *Mechanism.* The explanation procedure shows that the phenomenon can be understood by explicating the details of the mechanism, the process, which ultimately brings about the phenomenon in question. For example, the mechanism of how a flashlight generates light can be explained by determining its components (a battery, a light bulb, etc.) and describing how they are connected to one other, with each part being explainable by the analysis of the mechanism that operates it. The immediate recall of a seven-digit number can be explained by the explication of the cognitive mechanism based on the distinction between short-term memory (STM) and long-term memory (LTM).

## *Analogy.* The explanation procedure shows that the phenomenon can be understood by showing the similarity between the phenomenon being investigated and a different phenomenon that is understood. For example, the comparison between the human cognitive system and the operation of a computer. As mention above, human memory is generally explained by the explication of mechanisms for input, encoding, storage, and retrieval that correspond to computer mechanisms.

1. The explanation procedures (models) are based on logical and empirical justifications.

According to Hempel (1965, 1966), the fundamental justifications of the explanation models he proposed (such as the D-N model) are rational and empirical requirements (see discussion in Rakover, 2018). The rationality requirement is that the model or explanation procedure does not lead to internal contradictions or inconsistencies. The empirical requirement is that the explanation procedure is closely related to the studied phenomenon.

1. Two necessary conditions for understanding and explanation.

Understanding and explanation must fulfill two necessary conditions (see Rakover, 2018, in preparation). The first condition is that understanding of some content is impossible unless that content is represented (or was represented) in the consciousness of the person, i.e., consciousness is a necessary condition for understanding. (Consciousness is not a sufficient condition because many other conditions must be fulfilled for understanding to occur, such as normal functioning of the brain.) The consciousness condition is of great importance because it makes it possible to distinguish between an explanation and the understanding of an explanation. To illustrate this, let us look at the following example.

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 explanation as well as the worst of his students? **My answer to that question is no because a robot does not understand either the questions nor the answers that it provided its students. (In Rakover, 2018, I review the claim that even a sophisticated and complex robot has not developed anything similar to human consciousness.)**

The second condition for understanding an explanation is that understanding of content occurs when questions are asked about the content discussed and answers are provided that accord with the relevant knowledge of the content and the explanation procedures. For example, the answer to the above question about free-falling bodies accorded with the accepted knowledge of the period and followed the relevant explanation procedure, i.e., Hempel’s N-D model (Hempel, 1965, 1966).

1. The phenomenon under investigation cannot be understood if its explanation is not understood.

This requirement is 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 this 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.

This idea, which is at the core of the discussion in this article, will be explicated and explained by discussion of the problems related to the lack of understanding of complex computer programs the progressive AI models.

# 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 softwares, 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 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 softwares 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 softwares 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 (softwares) 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. 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

In this section, I want to clarify the methodological and philosophical implications associated with the problem of our lack of understanding of progressive AI models and of the programs that are meant to interpret them. First, I will 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, in a way similar to Robert the robot, one may propose that not only do progressive AI models not understand what they are doing, but even their explanatory programs do not understand what they are explaining, because all of these types of software lack consciousness (as mentioned above, Rakover, 2018, argued that yet there is no computer that 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 these softwares 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.

# A Progressive AI Model as a New Phenomenon and Different Levels of Understanding

The fact that progressive AI models are not understood inspires the production of explanatory programs (softwares) 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 as a new phenomena that need to be understood by explanatory programs?

**I will discuss two points. The first one points out that there is a difference between a natural phenomenon and an inapprehensible progressive AI Model as a new phenomenon. While the new phenomenon, inapprehensible progressive AI model, may be incorrect, a natural phenomenon is neutral in this respect. If the inapprehensible progressive AI Model is incorrect (an incorrect theory may generate correct predictions), then a good explanatory program will deceive a human user when it will not point out that the inapprehensible 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 explanatory softwares do not provide complete explanations to not-understood progressive AI models but only partial explanations. This raises the following point: despite the above difference in empirical testing between a scientific theory and an explanatory program, 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 method 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 even these AI models do 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: The Analogy to Cognitive Psychology

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 mechanism, 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 sub-systems 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 comparisons between humans and machines, see Borowski et al, 2021)**.

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 explanation for B to the not understood behavior A. However, when operation 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 the progressive AI models, the most successful in predicting behavior, it is possible to propose the critique argument suggested and supported here. Given this, one may take an extreme provocative position and wonder what is the motivation for continuing to build these kind of models for explaining behavior when these models themselves are incomprehensible? To answer this question, one needs to evaluate in a broad theoretical approach the main advantages and disadvantages of the computer analogy upon which cognitive psychology is based.**

**Advantages:**

A. Some researchers propose the following interesting analogy: just like it is possible to reduce every computer program to a series of zeros and ones (absence or presence of voltage), so it is possible to reduce all cognitive and mental processes that occur in a person’s mind to the neurophysiology of the brain. In other words, the analogy to a computer made possible the development of a tempting hypothesis that the solution to the mind-body problem will ultimately be found, i.e., it will become possible to develop a theory that explains how mentality-consciousness can be reduced to the neurophysiology of the brain (see Dennett, 1979; and critical discussion in Rakover 2018).

B. The computer analogy is not a general theory that addresses many different types of behavior like the general theories of Newton, Einstein, and quantum mechanics. The analogy is a general framework (like the rules of chess) within which one may construct a variety of hypotheses, theories, and models that relate to a variety of behavioral phenomena over a wide spectrum of types of behavior. That is to say, the analogy provides fertile ground for the growth of specific models and the discovery of many interesting behavioral phenomena. (Note, however, that several researchers consider the absence of a general psychological theory to be a disadvantage, see Rakover, 2020.)

C. **Cognitive psychology may suggest that incomprehensible progressive AI models do not offer zero understanding, but rather some understanding that is rooted in the algorithms used to create these models, and in the explanatory programs that somewhat clarify these models explanatorily.**

**Disadvantages:**

A. 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). At the extreme, this shift from emphasis on explanation to prediction results in the acceptance of any incomprehensible theory or progressive AI model, as long as it successfully predicts the observed results. This extreme 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. We will have a difficult time distinguishing between many possible theories that predict 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 stated above, without explanation and understanding scientific progress will be stopped and ruined. Yet, without predictions (to be compared with observations) science cannot progress either. So, I would propose that explanation (understanding) and prediction (observation) are both necessary conditions for scientific progress.

B. 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. it is not influenced at all by mental processes such as desire and belief, human (and animal) functioning 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 (**for arguments supporting this approach see** Rakover, 2018).

C. In the analogical transfer of mathematically well-defined concepts from computer science to the field of cognitive psychology, the boundaries of the basic concepts, ‘information’ and ‘information processing’ were breached. While these concepts are precisely defined in computer science, they are applied imprecisely in psychology. For example, ‘information’ can refer to almost anything, including syllables, words, sentences, excerpts of poetry or prose, visual images, etc. (see discussion in Palmer & Kimchi, 1986; Rakover, 2018).

**Given these advantages and disadvantages, we have to go back to our previous question and ask ourselves more generally: what are the implications of incomprehensible progressive AI models to cognitive psychology? The answer depends on the following two possibilities:**

1. If we think the advantages outweigh the disadvantages, we can continue with business as usual in cognitive psychology and be pleased that progressive AI models provide us with a certain limited and partial level of explanation of behavior.
2. If we conclude that the disadvantages outweigh the advantages, we need to replace the analogy of psychology with computers. If this is the case, then the following question must be asked: **Which scientific approach is a good candidate to replace the computer analogy?**

Since I do not have an answer to this last, difficult question, I will leave it unanswered and end the article here.

# References

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.

Dennett, Daniel C. (1979). *Brainstorms: Philosophical Essays on Mind and Psychology*. The MIT Press.

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

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, Carl. G. (1965). *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. The Free Press.

Hempel, Carl. 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

Palmer, S. E., & Kimchi, R. 1986. The Information Processing Approach to Cognition. In Terry J. Knapp & Lynn C. Robertson (Eds.), *Approaches to Cognition: Contrasts and Controversies*, (pp. 37-77). LEA.

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

Rakover, Sam S. (1990). *Metapsychology: Missing Links in Behavior, Mind and Science.* Paragon/Solomon.

Rakover, Sam S. (2018). *How to Explain Behavior: A Critical Review and New Approach.* Lexington Books.

Rakover, Sam S. (2020). Why has the field of psychology not developed like the

natural sciences? *The Journal of Mind and Behavior*, 41, 247-266.

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

Rakover Sam S. (in preparation). Two Factor Theory of Understanding

(TFTU): Consciousness and procedures.

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.*

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.

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.