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

**Author’s Name, Author’s Affiliation**

**Author Note**

**Author name ORCID ID**

**Disclosures/ acknowledgements**

**Contact information**

# Abstract

This essay discusses deep AI computer models that have been created in order to understand human behavior. 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 they were created to explain. 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 deep AI models, the most successful and sophisticated programs that are meant to explain behavior, it is not clear why we should continue building such computational models, since they do not serve to help us to understand behavior. This article discusses this question while briefly describing the basic principles of explanation models, the problem of our lack of understanding of deep AI models, 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**

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: reception, coding, storage, and information retrieval. This analogy also offers a way to solve the body-mind problem: some researchers believe that the grounds of mental processes can be found in the neurophysiology of the brain in a way that is analogous to the way software in a computer is ultimately physically grounded in electrical current (there either is or is not an electrical charge. See discussion in, for example, Dennett, 1979; Rakover, 2018).

The problem that arises here is that the most advanced models of AI for understanding the behavior of the individual, deep AI models, are so sophisticated and complex that we cannot understand how they operate; they should actually be 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 explain. This is a very strange situation because it 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 this whole psychological approach. (In this article, I will use ‘deep AI models’ to refer to a large variety of different types of software such as machine learning and neural networks; the important differences between them are not relevant to my argument).

This article includes a brief introduction that summarizes the basic assumptions of an explanation model, a description of the problem with our lack of understanding of deep AI models, 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 unexplained 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 an unexplained phenomenon, one must use an explanation model: a special procedure that is appropriate to the type of phenomenon to be explained.*

This requirement is based on the fact that the phenomenon has been observed to meet 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 an attempted explanation of this representation is expressed as 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 the law of gravity (there are, of course, other procedures and 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 (see discussion in Rakover, 2018; Salmon, 1984, 1990):

## 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 with the help of 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 listing its parts (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 a 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 and known. For example, the comparison between the human cognitive system and the operation of a computer. Likewise, human memory is generally explained by the explication of mechanisms for input, encoding, storage, and retrieval that correspond to parallel computer mechanisms.

1. *The models, the explanation procedures, 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 (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 model or explanation procedure be closely related to the explained phenomenon. Even if the explanation is logically perfect (i.e., without any internal contradictions), if its concepts are not sufficiently related to the empirical phenomenon being investigated, the explanation will have no value. For this reason, among other things, it would be very difficult to use Newtonian theory to explain facial perception in humans.

1. *Two necessary conditions for understanding an explanation.*

Understanding an 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 two scenarios.

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 or the answers that it provided its students. (In Rakover, 2018, I present a review of the evidence for the claim that even a sophisticated and complex robot has not developed anything similar to human consciousness.)

Here is another example that testifies to the fact that however complex and sophisticated a robot may be, it will not develop consciousness: Suppose a robot mother perfectly mimics the behavior of a human mother. From the moment of receiving the newborn, this robot mother acts exactly like a human. Just like a human mother, this robot mother would reject a robot child in favor of a human child. Here is the key point: If a robot mother had indeed developed consciousness like a human, it is reasonable to suppose that she/it would reject a human child and embrace a robot child who is much more readily conceived of as her/its offspring. However, she/it does not behave like this, because she/it is nothing but a machine that has learned to respond with a particular pattern of reactions when presented with a certain stimulus pattern.

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 knowledge of the content and the relevant procedures for giving explanations. 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 either wholly or even partially successful.

This idea, which is at the core of the discussion in this article, can be explicated by discussion of the problems related to the lack of understanding of complex computer programs that use artificial intelligence (deep AI models).

# The Lack of Understanding of Deep AI Models

The cases that I will address in this discussion are related to deep AI models, including different types of sophisticated and complex software that are used to explain, among other things, human memory, facial recognition and identification, decision making, and categorization, including medical diagnoses. (See, for example, Elmahmudi & Ugail, 2019; Kumar, A., 2021; Samek, Montavon, et al, 2019; Samek, & Muller, 2019; Samek, Wiegand, et al., 2017; Zhou, Bau, et al., 2019). This type of software, deep 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).

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

Despite the revolutionary charter 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 deep AI models. 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 only partial facial information (as opposed to not training in this way). In this case, one may also claim 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 training 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.) 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 Samek, & Muller, 2019). Although explanatory software does help to understand the deep AI models, the same disturbing question arises: Do we understand the explanatory software? 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:

More recent deep learning based neural networks provide far superior predictive power, but at the price of behaving as a ‘black box’ where the underlying reasoning is much more difficult to extract. (p. v).

They go to question the explanatory software that is designed to provide an understanding of deep AI models:

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).

In this article, I will address both deep AI models and explanatory software that are not understood; i.e., I will address all the sophisticated and complex programs that are considered black boxes.

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

I do not want to dwell on explainable AI software in this article, even though it is an important topic for investigation. Instead, I want to clarify the methodological and philosophical implications associated with the problem of our lack of understanding of the explanation provided by deep 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 relate to software that is not understood.

# 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 describes precisely 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, not only do deep AI models not understand what they are doing, but even their explanatory programs do not understand what they are explaining (i.e. the functions they are running) because all of these types of software lack consciousness, just like Robert the robot. Second, while human beings can understand how Robert works, human beings, even the programmers of the software themselves, cannot understand the very complex actions performed by deep AI models or their 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 complexity of the deep AI models. Second, one can conceive of these programs as the broad frameworks within which events that require explanation take place. The principles by which the deep AI models were designed are insufficient to explain these events. This idea can be explicated by the analogy to chess (or any other game).

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 Bobby Fischer was one of the greatest chess players just by explaining the rules. To understand how Fischer 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 deep 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 deep AI models are no more than the rules that set up the framework within which a program will develop that is so complex 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 deep AI model is a ‘black box.’ It is for this reason that we need explanatory programs to explicate these opaque models.

# A Possibility for Understanding an Incomprehensible Program: The Deep AI Model as a New Phenomenon and Different Levels of Understanding

The fact that deep AI models are not understood inspires the production of explanatory software as well as studies that use experiments to decipher what they are doing (see, for example, Elmahmudi & Ugail, 2019; Samek, Montavon, et al. 2019). This raises the possibility that in producing their results, these programs generated new phenomena that need to be explained, i.e., the programs themselves have become objects of interest that we seek to understand. Given the fact that deep AI models, including their explanatory software, are not understood, the following question arises: how are we to relate to the fact that these programs are not understood and should be regarded as new phenomena that need to be understood themselves?

One can regard deep AI models and their explanatory software as providing partial, imperfect explanations, i.e., as providing various low levels of understanding (of either the object of study or of the not understood program used in its study). 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. From this data, it is clear that the explanatory software does not provide complete explanations of deep AI models and in this regard, it is not different from other modes of scientific explanation that are characteristically partial. 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 principles do not apply when a body’s velocity approaches the speed of light.

1. The Scope of the Explanation:

 Every empirical theory is limited, either explicitly or implicitly, by some empirical or theoretical boundary. The theory fully applies to all relevant phenomena within that scope. For example, Newtonian mechanics applies to all bodies moving at normal 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 for facial recognition, “the Catch model,” whose principles set up theoretical-empirical constraints on an individual’s behavior (the witness). With the help of this model, the researchers attempted to reconstruct the target face from the witness’s memory. One of the most difficult problems with this method arose from the fact that it did not take into account the fact that exposure of the witness to additional faces beyond the target face (exposing the witness to additional faces was part of the model’s method of facial reconstruction) interfered with the individual’s memory and significantly limited the usefulness of the model. In other words, if a person’s memory of faces was not negatively affected by additional facial information, it is reasonable to conclude that the Catch Model would be more successful at reconstructing the target face than the 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 can conceive of deep AI models as so complex that they are not understandable; they are black boxes. On the other hand, these models do provide a certain, limited level of understanding that is derived from the way they were programmed. For example, one can achieve a shallow understanding of how certain neural networks are developed and trained using a special algorithm called backpropagation: in short, this algorithm uses a mistake made by the software (the gap between the output value and the normal behavioral value) to change the weightings (the strength 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 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 our conjectures about certain parallel 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.

The similarity between 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 useful or correct explanation is the fact that every data set can, in principle, be derived from an infinite number of different functions (i.e. theories). In this case, the data involving the state of affairs that is the impetus for the signing and the response of signing or the state of affairs that is the impetus for the pouring of the tea and the response of pouring the tea are 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-operation 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 most successful and complex programs, the deep AI models, what is the motivation for continuing to build models to explain behavior that are based on hidden computational processes? To answer this question, I will now analyze some advantages and disadvantages of the analogy upon which cognitive psychology is based.

*Advantages*:

A. As I pointed out above, some researchers believe that just like it is possible to reduce every computer program to a series of zero’s and one’s (or 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 an interesting and 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 critical discussion about this in Rakover 2018).

B. This 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.

C. Even if we conceive of the brain as some sort of natural super-computer and even if we understand the operation of most deep AI models that are designed to explain human behavior, these computer programs still provide a certain (low) level of understanding as I discussed above.

*Disadvantages:*

A. Our lack of understanding of deep AI models is the source of a serious methodological flaw. The present state of affairs is that deep AI models represent a high level of cognitive psychology, yet they are not capable of providing us with a reasonable 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 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, on understanding, to prediction results in the acceptance of any theory, 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 testing models and theories. Since there are countless functions that can produce good predictive data, absent a theoretical explanation that we understand, we will have a difficult time distinguishing between the many possible theories that predict the observed results. In fact, these well-predicting theories will become ad hoc, since it will be impossible to evaluate them. This argument is founded on the common-sense notion that the correct explanation produces successful predictions (i.e., that under the relevant conditions, the correct theoretical explanation will propose a successful prediction that will be realized or not realized). However, successful prediction is only a necessary condition of a correct explanation. A correct explanation cannot produce false predictions, but an incorrect theory can produce correct predictions.

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 permeate a great deal of behavior. As long as we do not have a theory that explains how consciousness is grounded in neurophysiology, it is difficult to see how deep 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 (see a review of this topic in 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 equivocally in psychology. For example, ‘information’ can refer to almost anything, including syllables, words, sentences, selections of poetry or prose, visual images, etc. (see discussion in Palmer & Kimchi, 1986; Rakover, 2018).

Given these advantages and disadvantages, we must ask ourselves how we should relate to cognitive psychology. The answer depends on the following possibilities:

1. If we think the advantages outweigh the disadvantages, we can continue with business as usual in cognitive psychology since deep AI models provide us with a certain (low) level of explanation and in some cases, explanatory software can be helpful in understanding these complex models.
2. If we conclude that the disadvantages outweigh the advantages, we need to replace the analogy to computers. If this is the case, then the question must be asked: What approach is a good candidate to replace cognitive psychology?

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

# References

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

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

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.

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 (in preparation)

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.

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.

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.*

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.