International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

On the Knowledge of Neural Networks: The Genesis, Characterization, and Justification

Huikang Jiang

Researcher, Beijing National Day School

Abstract

This paper systematically analyzes the knowledge generation process of neural networks through the lens of Philosophy for Science and Epistemology. This paper first argues that for neural network's knowledge to be possible, it is necessary that the truth of the world can be represented by mathematics, since neural networks are, by their first nature, a mathematical model, and this paper provides few arguments that support the mathematical nature of truth. For a central characterization of neural network's knowledge, this paper argues that neural networks bypass the three inherent limits of traditional sciences, such as physics and chemistry, suggested by Eugene P. Wigner (1995). The three limits are: 1) Traditional sciences are inevitably approximations of the transcendental truth; 2) Scientists will never stop pursuing deeper, more profound scientific theories that encompass more information than previous theories, despite every theory being an approximation; 3) The increasing complexity and depth of the successive scientific theories poses a significant challenge to human intellect. Neural networks bypass these limits by adaptively learning any underlying function to the given data space, without creating any fundamental hypothesis, which are inevitably not transcendental, like that of traditional sciences. This is also another distinction that supports the differentiation between "machine knowledge" and "human knowledge", as first proposed by Wheller (2016). This paper argues that the only way humans can justify machine knowledge is through Reliabilism, the epistemic justification through trust in the knowledge generation process. Although reliabilism may currently seem unacceptable, this paper predicts that scientists will take control over machine knowledge in the future, as testified by scientists exploring the chemical world, which is similarly beyond direct comprehension. In the end, this paper suggests the necessity of discussing how AI can justify its own knowledge as conscious beings.

Keywords: Epistemology, Philosophy for Science, Artificial Intelligence, Neural Networks

1. Introduction

In the past two decades, has Artificial Intelligence (AI) become successful and started to impact every aspect of people's lives worldwide (Makridakis, 2017). While scientists worldwide are using AI techniques as an efficient and effective instrument, the mechanisms through which AI models and training algorithms generate knowledge and predictions are generally unclear (Doshi-Velez & Kim, 2017). That is, foundational topics such as model parameter interpretations, the predictions' causal explanations, or the selected algorithms' reliability within a specific context are unanswered. Hence, AI models are often referred to as "black boxes." However, a better understanding of the foundations of AI is urgently needed since the use of these models has spread throughout the fields of facial recognition, self-driving cars, and medical diagnosis, where safety and morality are at stake, and there have been more than a few cases of

AI failure (Yampolskiy & Spellchecker, 2016). At present, the philosophical discussions on artificial intelligence are centered on its ethical issues and social impacts. At the same time, relatively little research exists on the philosophy behind AI-generated knowledge. To address this issue, I turn to epistemology, the subfield of philosophy that discusses the characteristics of knowledge and the methods of this acquisition. This paper attempts to epistemologically analyze the knowledge generating process of AI and possible justification for the generated knowledge.

Such analyses inevitably involve many types of models and variations of AI, each worth its examination. This paper will only investigate neural networks since they are the dominant models of today's AI. Neural networks' application is seen in almost every artificial intelligence subfield, including image classification, speech recognition, and natural language processing, as shown by multiple surveys on the field (e.g. Abiodun et al., 2018). The neural networks are unique AI models since they are particularly adept at learning high-dimension, non-linear, complex relations. Especially their unlimited ability to build low-level features up to higher dimensions and to fit any form of function, as they are Universal Approximators, marks a clear difference from other models. While no current literature particularly examines neural networks, the broader analysis of machine learning and artificial intelligence can provide a general framework for reference since neural networks are no more than a subset of these fields. This allows us to infer the potential perspectives and understandings the academic community might have regarding neural networks. For example, Frické (2014) is a critical analysis of Data Science methods for the irresolvable inductive biases; Smart et al. (2020) argues that Epistemic justification to machine generated knowledge is not enough and there ought to be moral justification; Hammoudeh et al. (2021) examines the common belief that more data leads to better machine learning models and the limits of using generic mathematical theories to describe the learning process; Wheeler (2016) and Bai (2022) develop the idea of "machine knowledge" as opposed to "human knowledge," and arguing that humans are unable to comprehend "machine knowledge."

The current discourse about machine learning and artificial intelligence epistemology in academia and the press has generally been critical. Researchers acknowledge that they can collect a large amount of data; however, selecting the categories to be collected is conjectural, and people need to learn what type of data can inherently offer explanations, theories, or solutions to scientific problems. Data is collected through instruments, which is fallible. Commentators criticize that machine learning and artificial intelligence's inductive nature may fail to provide valid theories.

Such tasks as fitting machines, machine learning, and artificial intelligence may be targeted with a few criticisms. First, the causal relationship is unknown. Similarly, people do not know whether there is a law-like relationship, such as in social sciences, where laws can be elusive. Second, there can be omitted variables. For example, a statistically significant number of smokers get lung cancer and also get cirrhosis in the liver. In this case, smoking causes lung cancer but not cirrhosis; it is just that many smokers happen to drink. Plotting the relationship between smoking and cirrhosis may have accurate results, but it cannot indicate any causation, and the factor that smokers often drink is omitted. Third, there can be inductive bias; that is, the formulation of the curve, whether quadratic or exponential, is unknown. Fourth there are problems of overfit and underfit.

Critical as these prior analyses may be, a few common issues remain unanswered within them. First, they take the statement "artificial intelligence is a knowledge generating enterprise" as an assumption without explaining the genesis of machine knowledge. This statement is instinctively true since there is such a broad success in applying artificial intelligence, testifying to its effectiveness. As a result, the part of the

discussion on why and how AI is a knowledge generating enterprise is often left out. This question should be further investigated because no one can understand the characteristics of knowledge generated without knowing anything about the generation process. Completing this part of the discussion will guide future discussions involving machine learning's epistemology. Second, the previous discussion focused on how data science/machine learning/artificial intelligence serves as a science. As a type of science, the sole problem in Epistemology is conveying the model's results for human use, with the concern of human safety and understanding as a central purpose. However, scientists are building AI to mimic human intelligence and are concerned about intelligence crises, such as emergent awareness and AI taking over the human world. Therefore, another necessary field of discussion is how AI will justify its knowledge, assuming it gains intelligence and self-awareness. In other words, we need to consider AI's epistemology from the perspective of the emergence of intelligence, AI as the subject themselves. Third, the recent discoveries of emergent world models in Large Language Models shed new light on the topic (Gurnee & Tegmark, 2024). For the first time, scientists successfully detected one internal structure of Large Language Models. This discovery should bring researchers new insights into how people take machine-generated knowledge.

This paper will analyze how neural networks generate knowledge and present an optimistic view of machine knowledge while addressing the three unresolved issues mentioned above. Although there are variations to the definition of knowledge, there is a consensus between most epistemologists that knowledge should be minimally given by a definition of three parts: justified and true belief (Ichikawa, 2020). I will adopt this definition to guide my analysis. This paper should discuss each of these three elements. The belief element is beyond the scope of the paper; here, I will say that neural networks "believe" anything encoded in its numerical parameters. For the "truth" element, which I discuss in section 2, I must consider what characteristics the truth of our universe holds and answer the question of why neural networks can fundamentally learn the truth. I argue that mathematics underlies all phenomena or can be mathematically represented. This trait of truth is necessary because neural networks are natural mathematical models, which means they cannot learn anything beyond the scope of mathematical representation.

Furthermore, as Universal Approximators (Hornik et al., 1989), neural networks can learn any designated function. That is, for whatever mathematical function governs the phenomena, neural networks can learn it. This is why neural networks can learn; the process they learn is their training process. In section 3, having the philosophical explanation of neural networks generating knowledge, I move on to develop the idea of "machine knowledge" by providing a new distinction with "human knowledge", namely, neural networks are capable of overcoming the limits of "layers of science," a concept proposed by (Wigner, 1995). According to Wigner, physics, chemistry, social sciences, and other forms of traditional sciences, such as those before data science and machine learning, are inevitably constructed by layers and layers of theories. Each theory is a human construction that models the target problem, with new theories encompassing more phenomena than older ones. The progression from Newtonian Mechanics to Relativity is as such. Each theory is inevitably an approximation to the problem. However, neural networks do not fall for these limits because they can accurately model the target mathematical function, which I will further explain below. Since the layer structure is inherent in all traditional science but not in any neural networks, this difference distinguishes machine knowledge and human knowledge, and neural networks also surpass human knowledge in this perspective, since they do not fall for this limit.

Finally, in section 5, I will consider the possible epistemic justifications for machine knowledge. As machine knowledge is inscrutable, this indicates only one possible form of justification: Reliabilism, justification by the trust in the process of knowledge generation. While such a justification may seem unfulfilling, I bring up examples such as the Linear probe to demonstrate that researchers would find methods to understand machine knowledge using machine knowledge. In the end, from considering AI as an intelligence subject and how it justifies its knowledge, I provide the discovery of emergent world models as proof that such a discussion is necessary.

2. Truth

Before discussing whether neural networks are learning knowledge, I must first discuss the properties of the truth of our world. *Belief* adheres to truth, and thus, the properties of the truth will subsequently influence the knowledge production of neural networks we care about. For example, if there is no single truth, one may argue that neural networks may never learn anything. Similarly, other theories on truth will lead to other conclusions.

The first and foremost characteristic in this analysis is the effectiveness of mathematics in modeling the universe or mathematical functions that can represent the problems researchers are trying to solve. Neural networks are, by their very nature, a mathematical model. Each neuron in the network performs a mathematical operation, typically a weighted sum of its inputs followed by an activation function. This allows the overall network to be seen as a complex mathematical function that maps input data to output predictions. The neural network parameters, including weights and biases, are numerical values adjusted during training using optimization algorithms like gradient descent, which are rooted in mathematical principles. The training process involves minimizing a loss function that quantifies the difference between the network's predictions and actual target values, employing calculus-based methods to find the optimal set of weights and biases. Many scientists have developed more complicated variations of neural networks, such as Bayesian neural networks and Graph neural networks. However, their variations inevitably incorporate other mathematical structures and nothing more. Regarding application, neural networks formulate a mathematical relation that maps the input data to the output. In essence, neural networks are in every aspect mathematical.

Before applying neural networks to resolve a problem, the premise is that the problem can be mathematically formulated. Today, when there is such a wide spread of good applications of neural networks, we must ask ourselves why our problems can be mathematically formulated. However, this question, which touches on the very nature of science and math, is a generally unresolved philosophical issue. In the following paragraphs, I will review the primary theories on this problem. Generally, this discussion should be separated into two parts: exact science and inexact science. This categorization between science was proposed by Auguste Comte in his seminal book Course of Positive Philosophy (Comte, 1876) and later developed by philosophers including Michael Polanyi and Stephen Toulmin. This paper defines *exact science* as the field of study based on measurements, empirical evidence, and quantifiable data in which theories and models can be tested with high accuracy and predictability.

In contrast, inexact science deals with subjects that are more complex, variable, and less amenable to precise measurement and prediction. These fields often involve human behavior, social systems, or natural phenomena influenced by many variables, many of which cannot be controlled or isolated. Generally speaking, exact science accounts for mathematics, physics, chemistry, and biology, subjects known to have formal representations; inexact science accounts for sociology, humanity, and linguistics. The

problem of natural language processing, where humans have struggled to find sufficient formal representation, is categorized as inexact science. Since the two categories are treated with different philosophical analyses, I will discuss the two categories of exact and inexact science separately below.

2.1 Exact Science

In Eugene Paul Wigner's seminal essay, "The Unreasonable Effectiveness of Mathematics" in the Natural Science (Wigner, 1990), he proposed the problem that "the enormous usefulness of mathematics in the natural sciences is something bordering on the mysterious and that there is no rational explanation for it." Later, it was mentioned that "...the mathematical formulation of the physicist's often crude experience leads in an uncanny number of cases to an amazingly accurate description of a large class of phenomena." Wigner was the Nobel Laureate of Physics in 1963, and when he proposed this problem, his central concern was physics and other natural sciences, which we categorize as exact science. Bochner subsequently expressed a similar view: "what makes mathematics so effective when it enters science is a mystery of mysteries, and the present book wants to achieve no more than explicate how deep this mystery is" (Neugebauer, 1969). This question initiated a discussion to explore the mysterious relation between scientific theories and their mathematical formulation to formulate an answer.

Among many scholars answering this question, a common school of thought is Realism. At the heart of the Realist perspective is the belief that mathematical entities and structures exist independently of human thought and are discovered rather than invented. This school of thought posits that mathematics is an intrinsic part of the universe's fabric. The British mathematician Hardy expresses this view: "I have myself always thought of a mathematician in the first instance as an observer, who gazes at a distant range of mountains and writes down his observations" (Rose, 1988)

The Realists have plenty of evidence within the history of physics that indicates mathematics' inherency in physics. The Dirac equation in quantum physics, for example, not only described the electron but also predicted the existence of the positron before it was discovered. Dirac said, "It was found that this equation gave the particle a spin of half a quantum. Moreover, it gave it a magnetic moment. It gave just the properties that one needed for an electron. That was an unexpected bonus for me, completely unexpected" (Dirac, 1977). Dirac published his equation in 1927, which predicted the existence of a spin of half a quantum and remained undiscovered until 1932. This predictive power indicates that mathematics is not just a tool but a fundamental aspect of the physical universe.

Another example is the wide application of Maxwell's equations. The equations can accurately describe various electromagnetic phenomena, from transformers to radio transmission, showing that the mathematical structure has a rich ability to describe natural phenomena, which manifests the likelihood that mathematics underlies natural phenomena. Furthermore, the Maxwell equations also predicted phenomena that still need to be discovered, such as the existence of radio waves. Maxwell published his equations in 1864, whereas radio waves were not discovered until 1888. This predictive ability of the mathematical structure for unknown phenomena further supports the realist philosophy.

Experience has demonstrated the power of mathematics in modeling the problems of exact sciences. Mathematics has successfully represented problems in exact science and provided predictions for undiscovered phenomena. Therefore, it has been accepted that mathematics underlies the events in exact science and is transcendental to this knowledge.

2.2 Inexact Science

Not much literature delves into the relationship between mathematics and inexact science. As far as Google Scholar provides, the sole text that discusses this issue is On the Epistemology of Inexact Science

(Helmer & Rescher, 1959), where the authors Olaf Helmer and Nicholas Rescher proposed several reasons why mathematics can be incorporated into inexact science.First, exact parts exist within the inexact science, as they write, “Some branches of social science (e.g., certain parts of demography), which are usually characterized by the presence of a formalized mathematical theory, are methodologically analogous to the exact parts of physics."The second key aspect is the role of probability theory and statistics, which provide a mathematical framework to handle uncertainties and variations within these sciences. The authors state, "Mathematics offers the tools to quantify and model the probabilistic nature of inexact sciences." This probabilistic approach allows for creating models that can predict outcomes with a certain degree of confidence, even if they cannot offer absolute certainty. Third, the iterative process of refining models is highlighted as crucial. Helmer and Rescher note, "Through successive approximations and the continuous incorporation of new data, mathematical models in inexact sciences are progressively refined." This iterative refinement ensures that models become more accurate over time, aligning closely with empirical observations. Therefore, Helmer and Rescher note the use of mathematics within inexact science. It is possible that mathematics is as effective in inexact science as it is in exact science.

2.3 A Review

Acknowledging that mathematics is inherent in scientific problems is crucial so that a mathematical function can represent the solution. If true, neural networks can function in the way people expect them to. A mathematical function that governs the given data space must exist for a Neural Network to have a target to adhere to through its learning process. If this is not true, neural networks can only be left to malfunction since they cannot learn anything besides mathematical functions. On the other hand, as long as the problem is governed by mathematics, neural networks will function because they are naturally Universal Approximators, meaning that they will fit any function in the given data space.

This analysis explains why and how neural networks derive knowledge. Since our world can be mathematically represented, and neural networks are entities that adapt to fit any given mathematical function, neural networks can produce knowledge of the world through data. The process through which they do so is the process of regression, and naturally, regression is the method of neural networks deriving knowledge.

In the above two sections, I have demonstrated why mathematics will likely represent exact and inexact sciences. However, although most philosophers are inclined to believe that mathematics has this property to some extent, this topic still needs to be solved with a final answer. In the Constructivist view, mathematics is constructed by the human mind and is not independent (Hersh, 1997). Constructivists would say mathematicians use vectors to describe speed not because vectors are inherent in speed but because vectors seem appropriate to the mathematicians’ minds. In some interpretations of the Platonian view, mathematics is more fundamental than all other disciplines of knowledge. However, it is still less fundamental than the Platonian “world of ideals” that mathematics is between the phenomenal and the true transcendental.

Nevertheless, I will take this property of mathematics as an assumption in my analysis because without it, there is no basis for the knowledge generated by neural networks to exist. This property of mathematics is the necessary condition for Neural Network knowledge, and as long as the neural networks are given the appropriate training, the sufficient condition.

3. Machine Knowledge

At this point, we have in hand an analysis of a foundational principle that gives the possibility of neural network's knowledge. What then, does this principle tell us about the characteristics of its knowledge? Or, in other words, before the discussion on the possible justification to neural network's knowledge, what can we know about the knowledge so that there is a cornerstone for the analysis on justification? Here, I make a fundamental differentiation between "machine knowledge" and "human knowledge", having in mind that this differentiation will be decisive to the possible justifications.

The concept of "machine knowledge" was first put forward by (Wheeler, 2016) and later developed by (Bai, 2022). As Bai suggests, "the kind of epistemology or epistemic view held determines what kind of paradigm machine learning is designed." Our understanding of machine knowledge's characteristics will influence how we treat it. In the coming section, this paper will determine the justification for machine knowledge. In this section, I will review how Bai argues for the essential difference between machine and human knowledge. Then, I will develop this idea by adding one more crucial distinction between "machine knowledge" and "human knowledge." That is, machine knowledge can overcome a limit of human perception called "layers of science."

3.1 A Review of Bai, 2021

In his paper, Bai argues for a "machine-centric" epistemology and characterizes "Machine Knowledge" as an entity fundamentally different from "Human Knowledge," i.e., our knowledge. First, machine "experience" is characterized by data (large amount, multidimensional, rapid, automatic), which is different from human experience, and the assimilation of their experiences occurs via neural networks, which is an extension of the human perceptual system. The core difference is that machine knowledge relies on data, while human knowledge relies on information. Information is the observable representation of things, including much more detail than the data that machines use. Knowledge is formed when information is properly processed and combined with experience, judgment, and intuition——a process in which the human being as an epistemic subject plays a crucial role. In contrast, machine knowledge is primarily the recognition of patterns in data, which can be beyond human understanding and perception.

The relationships between the data that constitute machine knowledge are often beyond human understanding. Human perception is limited to three-dimensional physical space and one-dimensional time, and we can only partially perceive external information. The complex relationships that large-scale data machines can handle are not expressible through mathematical tools accessible to humans. As the layers and numbers of artificial neural networks increase, machine learning can handle increasingly complex data, leading to machine knowledge that is incomprehensible to humans. These observations align with the common critics posted towards machine knowledge, as briefly summarized in the introduction. Humans will not be able to perceive machine knowledge directly.

Machine "experience" is characterized by data, including large amounts, multidimensionality, rapidity, and automation. Machines can emulate aspects of human experience by calculating, computing, and correlating data. However, the precise mechanisms by which machine learning generates knowledge are often opaque and incomprehensible to humans, leading to the "black box" problem. This epistemic opacity of machine learning marks a difference to human knowledge, which is assumed to be understandable and transparent.

In summary, the author suggests that machine knowledge fundamentally differs from human knowledge, as it is based on data patterns rather than the richer information and subjective experience that characterizes human understanding. Therefore, Bai recognizes machine knowledge as different from human knowledge,

and thus, it must be treated as something naturally different from human knowledge.

3.2 The meaning of "layers of science" and limits of science

The concept of "layers of science" was proposed by E. P. Wigner in his famous paper "The Limits of Science" (Wigner, 1995). This concept is best illustrated with an example. Traditional Newtonian mechanics assumes that space and time are independent of each other and the observer's state of motion, and the interactions between masses are caused by force, as defined by Newton's Second Law of Motion, F=ma. For roughly two centuries after Newton published his work, Newtonian mechanics was seen as a complete and accurate description of the physical world, which was similar to people's attitudes to Aristotle's philosophy that objects fall at a speed proportional to their weight earlier; however, both theories do not represent the absolute truth. Although Newtonian theory does provide a great approximation for everyday situations, later scientists have recognized its limitations to macroscopic phenomena, such as the problem of blackbody radiation and the Michelson-Morley Experiment. In General Relativity, the interactions of masses are no longer derived from Force but are instead a manifestation of the Curvature of spacetime caused by mass and energy. In this case, the theory of relativity explains all the same phenomena Newtonian mechanics explains. Thus, the theory of relativity encapsulates and builds on the knowledge of Newtonian mechanics. It relatively resembles a higher "layer" of knowledge than Newtonian mechanics because it covers everything in Newtonian mechanics and pushes it slightly further. The structure of scientific theories, which extends from physics to psychology, is the structure of all sciences humans possess. Having one theory over another, this structure is, therefore, called "the layers of science."

In the past two hundred years, science has shown progress from Newtonian mechanics to quantum mechanics, field theories, and the relativistic quantum theory, and these can be considered as a total of four layers. Scientists are never confident that someone's layer will be the very last layer. As history has proven, no matter whether it is Aristotle's theory or the Newtonian Theory, and no matter how much people once were deeply convinced in its applicability, later scientists will realize their approximate nature and limitations and create another layer to incorporate more phenomenon, given an adequate amount of time. Aristotle's theory was false; Newtonian Mechanics was found only to approximate daily physical phenomena; Relativity has been found hard to reconcile with Quantum Theory. Scientific progress, as Wigner puts it, "always involve[s] digging one layer deeper into the 'secrets of nature,' and involve[s] a longer series of concepts based on the previous ones, those that are thereby recognized as 'mere approximations.'" This leads to the first limit of science: all science theories are approximations.

Although inevitably approximations, it is never one of science's courses to end its explorations. Regardless of how hard it may be, science will always continue its explorations, attempting to go beyond the scope of current recognition, and it will inevitably find extremities that current theories cannot cover. Science shall not be satisfied with Newtonian mechanics despite it being adequate for most worldly use cases. When science reaches the tenth layer of theories or any other number of layers, scientists have no right to expect that they have reached the final layer; whatever we have on our hands will mostly be another "approximation," and the search for new theories will never end. This is the second limit of science: scientists will never stop pursuing more layers of science despite every layer being an approximation.

In the future, there will be an accumulated successive layer of science, each with increasing complexity and depth, making it difficult for new students in science to dig through every layer to do research at the frontier. As Wigner points out, it will require a much more elaborate and much longer study to arrive at an understanding of the roots of the last layer, when the layers have accumulated, since every new layer

is marked by its ability to encompass more phenomena. This requirement conflicts with the deemed limited lifespan of humans and the limited intellect each scientist has. It is easy to imagine a stage in which the new student will no longer be interested, or perhaps practical, in digging through the already accumulated layers to reach the frontier. This increasing complexity and depth of the successive layers of scientific concepts poses a significant challenge to human intellect. This marks the third limit of science.

3.3 Neural networks overcoming the limits

It seems rather counterintuitive when Wigner comments that our science is approximations since humans hold the power of reason, which should have given accurate, reliable results. At least, when people criticize data science for its inaccuracy, they assume that Reason, as a faculty of mind, should bring people closer to truth than empirical speculations on data. I would rather not attack this assumption. What, then, causes our science to be approximations? Certainly not the mathematical tools since they have been so well tested. Then, there is only one thing left that can go wrong: fundamental principles.

The fundamental principles of scientific theories are not transcendental; rather, they are hypothetical products of the scientists' conjectures that have been tested, at best extensively, to be coherent and applicable and to have the greatest explainability of the world. The property of approximations is given by that they are hypotheses and conjectures. In Kantian philosophy, "transcendental" refers to the knowledge or concepts that are a priori, meaning they are independent of experience, precede it, and exist independently from human perception (Stang, 2024). In simpler words, transcendental knowledge is true in-them-selves, aiming for universality and necessity, and it certainly is not an approximation. This is the type of knowledge that people wish to pursue, as opposed to the knowledge produced by data science. However, as Wigner has pointed out, our sciences are approximations; all of humanity's science today has failed this undertaking of transcendental knowledge because their fundamental principles are not transcendental. Concepts such as Force in Newtonian Mechanics and Curvature in Relativity synthesize the experiment results and induction from experience. These concepts are in themselves a representation of the entire, real, physical world. The Concept of Force, though assumed to be true in its theory, is overthrown and replaced by Curvature in Relativity. Although the Newtonian theory cannot be established without this concept, the theory of Relativity functions even better without it. Since Force is only necessary for one theory but not for others, it is rather apparent that it is not transcendental since it is not a property of things true in themselves that vary with the condition. The principles of other theories are alike.

Kant's transcendental analysis aims to identify the general principles of physics as "constitutive principles of the metaphysics of nature" (DiSalle, 2013). However, Kant's theory of space and time is a philosophical theory, not a physical one. From Kant's point of view, the transcendental perspective is precisely what is lacking in Newton's metaphysics. At the same time, Newton's principles in his theory play a "constitutive role" in his theoretical framework. Unlike Kant's transcendental principles, they are not proposed as "necessary and sufficient conditions for a general metaphysics of nature." Instead, Newton's principles are "justified, not by independent arguments from metaphysics, but by the constitutive role that they play in the conceptual system of physics——in other words, as 'conditions of the possibility of a practical account of nature,' not indeed of experience in general, but of physics as a coherent explanatory framework." Newton's principles are "relative to the physical theory of which they form a part, and their sufficiency depends on the theory's empirical sufficiency," unlike the wanted transcendental principles which aim for universality and necessity.

Like other mechanics of his time, Newton proposes his theory using a hypothetico-deductive method. He used his keen observations to observe and record the motions of physical objects. From the observations,

he proposed a theory that explains these motions. For some mechanics, all motions result from impacts, and for Newton, the three Newtonian Laws of Motion. The theory is then measured by its eligibility and practical usage, whether it forms self-contradictions, can explain the known phenomena, and whether people could build the entire building of physics on top of it. This process is the creation of the foundational principles of our science. Therefore, the foundational principles of science are not transcendental in the Kantian sense but truly approximations. Even today, when scientists recognize the discrepancies of Newtonian Mechanics, people still use it daily. This testifies that people do not necessarily seek transcendental knowledge from scientific theories but justify them for their utility. The property of our sciences as approximations is inherent. Thus, science theories are deemed approximations because of their very roots, principles, and approximations.

Now, imagine a methodology for deriving knowledge that does not involve creating principles and hypotheses. That is neural networks. As well known, neural networks are Universal Approximators (Cybenko, 1989; Hornik, 1989). This property, mathematically substantiated by the universal approximation theorem, asserts that a neural network with at least one hidden layer and a sufficient number of neurons can approximate any continuous function to any desired degree of accuracy, given appropriately trained weights and biases. Witness that neural networks are fundamentally the summation of a large quantity of non-linear activation functions, such as the sigmoid or ReLu, which empower neural networks to capture complex, non-linear relationships within the data and further enable the modeling of a wide array of phenomena. Each activation function can be tuned on its own parameters, based on the training from input and output data, to represent local aspects of the targeted data space, and these local approximations collectively form a global approximation of the target function. The central point is that the neural networks learn directly from the given data set and discover, or approximate, the hidden function. For example, if Newton's second law were absolutely, transcendentally true, and Force always equals mass times acceleration, a neural network would eventually learn this, given sufficient size and training time. However, unlike standard scientific theories, neural networks are very adaptable. They are not confined to theories that must stem from a hypothesis, nor are the forms of their theory confined when their hypothesis is settled. Continuing the example, we presently know that Newton's second law is not true at high speeds. Neural networks could learn this discrepancy if given access to data at high speeds. The neural networks will reach it no matter what the true mathematical equation governs the relationship between the input and output data. In another perspective, the mathematical representation of Newtonian mechanics is determined to be F=ma once the hypotheses are decided. The equation is an inevitable product of a series of logical deductions starting from the hypothesis. So is the mathematical representation of relativity determined when its hypotheses are settled. Neural networks do not fall for this limitation, they learn whatever mathematical representation there is. Whether the transcendental truth is that F=ma or some other formulation in another theory, the neural networks can learn it, since the transcendental representation is inherent in the data. The process of neural networks learning completely bypasses the limits of approximations, as it is not based on human-presumed empirical principles. Neural networks showing greater accuracy in complex tasks than human-built theories testify to neural networks' effectiveness.

Neural networks' advantage is exemplified clearly through its excellence in natural language processing. Rule-based systems, also known as symbolic AI, are constructed based on a predefined set of rules that encode domain knowledge in the form of logical "if-then" statements. This approach to AI can be considered more analogous to human science since the rule-based systems are constructed based on

empirical assumptions and developed into human-deduced theories. An input data is given, and the output is derived from the pre-designed set of deterministic rules, like a handbook. However, as the complexity of the domain increases, maintaining and scaling these rule sets becomes challenging. Symbolic AI struggles in tasks that involve high variability and intricate patterns, such as object detection in complex scenes with diverse backgrounds and lighting conditions. Neural machine translation (NMT) models, such as the sequence-to-sequence (Seq2Seq) model with attention mechanisms, have long resolved rule-based systems' discrepancies by incorporating Neural Network techniques. For instance, Google's Neural Machine Translation system (GNMT) (Wu et al., 2016) implemented a Seq2Seq model with attention, significantly improving translation quality by capturing complex linguistic patterns. Facebook's Fairseq translation system (Sutskever, 2014) utilizes transformer models, a variant of Seq2Seq, which excel in handling ambiguity and variability in language, as demonstrated by its ability to generate contextually appropriate translations with fluency and coherence. ChatGPT and similar LLMs are also based on transformer models. So, these neural network methods are much more effective at language than the rule-based methods analogous to human scientific endeavors.

3.4 Back to Machine Knowledge

Witness that the structure of "layers of science" is inherent to traditional science. Every traditional science is doomed to endure the limits derived from it. How humans structure their knowledge is represented by the way these traditional sciences do. These traditional sciences are said to be built by the faculty of reason and reflect how humans build their knowledge. As I have demonstrated, neural networks do not conform to this structure; therefore, "machine knowledge," derived by neural networks, is fundamentally different from "human knowledge." Since the "layers of science" come with disadvantages that neural networks can overcome, I hereby argue that neural networks can provide a better learning method than human learning.

4. Our Justification

In the above two sections, we have accounted for the properties of truth and the properties of the knowledge gained from neural networks. In this section, we move to the next step in the justified, true belief definition of knowledge and account for the possible justifications for machine knowledge. The first categorization we should make is between external justification and internal justification. Which of the two considerations of machine knowledge will greatly influence the result of our analysis? This distinction needs to be carefully handled in existing literature on machine knowledge.

In epistemology, the concepts of internal and external justification pertain to the basis on which a belief is deemed justified. According to the Stanford Encyclopedia of Philosophy (Pappas, 2013), internal justification posits that "a person either does or can have a form of access to the basis for knowledge or justified belief," which means that the reasons or evidence supporting the belief must be accessible or knowable by the person, emphasizing an introspective awareness of the justifying factors. The "basis for one's knowledge and justified belief" refers to the underlying reasons, evidence, or grounds that support or justify a person's belief or claim to know something. Simply, it means the subject's mental state and process. For a belief to be internally justified, the individual must be able to reflect upon and recognize the reasons or evidence that support it. The reflection often involves conscious awareness of one's mental states, perceptions, logical reasoning, or any other form of reliable evidence that ensures the belief accurately reflects reality. Internalism requires that these justifying factors be accessible to the individual's reflective awareness or introspection.

External justification, by contrast, denies that one must always have access to the basis for one's knowledge and justified belief, arguing instead that "things other than mental states operate as justifiers." This view allows for external factors, such as the reliability of the cognitive process that produced the belief, to serve as justifiers, regardless of the individual's awareness of them. Externalists maintain that a belief can be justified by objective factors like the dependability of the method by which it was formed, even if the person holding the belief is unaware of these external justifiers. Again, the primary difference lies in internalists requiring accessibility of justifying reasons, while externalists accept justification based on external, often objective, factors.

Now, we can clearly categorize previous analyses on the epistemology of data science and machine learning: they are inevitably external justifications. The scientists trying to interpret the inner mechanisms of the neural networks need access to the neural networks' mental state and mental process. In the above section, we have demonstrated the existence of Machine Epistemology, which we argued to be inherently different from human knowledge. Due to this difference, as humans, we can never directly perceive and understand the knowledge gained from neural networks. As humans, we can only partially understand the digits within the neural networks. Access to the mental state and process of neural networks is necessary for scientists to truly have internal justifications for the knowledge gained by neural networks. Alternatively, from another aspect, the process through which neural networks gain their knowledge is not something the scientist can reflect on using their introspective awareness.

In the previous literature, scientists are inherently considering external justifications. When addressing the fallibility of instruments used to collect data, which suggests a concern with the reliability of the processes by which data is gathered, scientists are considering the process and instruments through which the knowledge is generated, which aligns with external justification. When addressing the inductive biases of omitting variables, scientists examine objective factors that justify a belief, such as causal relationships and variables that lead to the result, which, again, is aligned with external justification. In fact, as a third person observing the learning of neural networks, we can only have external justifications, which is very intuitive since we are, in fact, external.

In the following two sections, we will address the issues of external justification and internal justification separately.

4.1 External Justification

In this section, we argue that the existing methods of justification for machine knowledge are reliabilistic, and we will predict the trend of the development of Reliabilist approaches, saying that in the future, we will have a coherent and sufficient method of external justification. For now, allow me to stop for a minute to conduct a brief overview of major kinds of justification defended in the philosophical literature.

Foundationalism posits that some beliefs are basic or foundational and do not require further justification from other beliefs (Pasillos, 2017). These basic beliefs serve as the starting point for all other justified beliefs, where all other knowledge can be derived through a set of logical deductions. The famous assertion “I think, therefore I am” and the classical syllogism are examples of foundationalist justification.

Coherentism argues that the justification of a belief is a matter of its fitting coherently with a system of other beliefs (Quine & Ullian, 1978). In Coherentism, the justification of a belief is not grounded in its correspondence to reality (as in foundationalism) or its inferential relationship to other beliefs (as in Evidentialism) but rather in its consistency and coherence within a broader network of beliefs.

Reliabilism argues that belief is justified if a reliable process produces it. Alvin Goldman is a prominent philosopher associated with reliabilism. There are two main types of reliabilism: process reliabilism,

which focuses on the reliability of the process that produces the belief, and indicator reliabilism, which focuses on the reliability of the indicators or evidence that supports the belief.

Imagine an experienced ornithologist and a novice walking through the forest. Suddenly, a Pink-spotted Flycatcher lands on a branch. Both observers think it is a Pink-spotted Flycatcher, and their judgment is correct. However, according to reliabilism, only the ornithologist's judgment is justified. This is because the novice is merely guessing, and his judgment needs a reliable source of information.

On the other hand, the ornithologist relies on his extensive knowledge and experience, matching his memory of birds with his visual observation of these birds. This matching process is reliable and based on accurate information and experience. Although the ornithologist cannot explain how he identifies birds in detail, his judgment is reliable because of his mass experience over the novice. This is an example proposed by Feldman (Pappas, 2023) that illustrates the essence of reliabilist justification, and this is how neural networks can also be justified.

For a well-trained bird classification algorithm with a massive amount of pictorial data as its training set, we understand that it is well-trained through the method we train it. Even if we cannot pinpoint which specific input features and correlations led the algorithm to identify the Pink-spotted Flycatcher, we consider its result to be mostly reliable within the data space. Using a machine learning algorithm for bird identification is dependable and can sometimes be more accurate than an expert birdwatcher. Although we cannot precisely explain the workings of the deep learning network, its reliability makes this explanation unnecessary.

The common methods we currently have of judging whether a neural network is well-trained fits in this category. For example, GPT-3 and BERT are transformer-based models used for various NLP tasks such as language translation, sentiment analysis, and text generation. Looking at the design of transformers, we can easily see that they allow the neural network to shift focus when analyzing a sentence. Through this intuition and practice, we vaguely know that transformers leverage self-attention mechanisms to understand the context and relationships within the text, which allows them to handle tasks that involve understanding language at a deeper level. Transformers' past success on related problems demonstrates its superior performance in capturing long-range dependencies and contextual information compared to traditional RNNs and LSTMs, which leads us to use it again. Although we may not be able to spell out why Transformers work in such a way, we use it frequently.

As we have shown, machine knowledge fundamentally differs from human knowledge, and it is easy to see the possible justifications we can have for neural networks. Reliabilism is the main possible justification we can have. Foundationalism requires certain foundational beliefs typically derived from immediate experiences or self-evident truths. Since foundationalism relies on direct perception or introspection to establish its foundational beliefs, it needs help to justify claims about the existence or nature of things that cannot be directly observed or understood. The leap from the concrete and perceivable to the abstract and inconceivable requires a form of justification beyond the immediate and self-evident, which foundationalism cannot provide. While the Coherentism approach may seem promising for justifying complex or abstract ideas, it risks creating a closed system where beliefs reinforce each other without necessarily connecting to the external world. For something beyond direct perception, the coherence of our beliefs does not guarantee their truth. The system could be internally consistent but still fail to represent reality accurately. For such an enterprise that is beyond direct perception and machine knowledge, the only choice we are left with is to believe in the reliability of the knowledge generating process.

We must further determine whether we should settle with Reliabilism or lose faith in AI. We should also determine whether a Reliabilist justification nature will impair our AI advancements. Most current analyses, like those stated in the literature review, reject this Reliabilist justification. People should have faith in AI because of the prospective field of explainable AI. There exist other instances where humans managed to understand a world beyond direct perception.

For the length of this paper, I will not cover all existing techniques for interpreting the inner mechanisms of neural networks, but I will present one solid example: the Linear Probe. A linear probe in the context of large language models (LLMs) is a technique used to analyze and understand the inner mechanisms of these models. It involves adding a linear classifier or layer with linear functions on top of a pre-trained neural network, such as a large language model. This probe is designed to be simple without introducing complex, non-linear transformations. The primary purpose of a linear probe is to determine how much information about a specific task is already captured by the pre-trained model. It helps in assessing the quality of the features learned by the model. To insert a linear probe into LLMs, researchers start with a pre-trained model trained on a vast amount of text data to learn general language representations. They then add a linear layer on this pre-trained model, which acts as the probe. The output from the last layer of the pre-trained model is fed into this linear probe. The linear probe is then trained on a downstream task, such as sentiment analysis or topic classification. Researchers use linear probes to understand LLMs in several ways. By training a linear probe on a task, they can assess how much of the task's information is already encoded in the pre-trained model's features. If the linear probe performs well, the pre-trained model has learned relevant features for the task. It also helps in understanding the transferability of knowledge from the pre-trained model to different tasks. If a linear probe is effective, it indicates that the model's representations are general and useful across various tasks.

The paper "Language Models Represent Space and Time" by Wes Gurnee and Max Tegmark delves into the internal workings of large language models by employing Linear Probes. This approach involves training a simple linear model on the activations of a neural network to infer the presence of specific features. In their study, the researchers constructed six datasets, each containing the names of places or events and their corresponding spatial or temporal coordinates. These datasets span various scales, from global locations to specific periods. They then used the Llama-2 family of LLMS, which are large transformer-based language models, to process the names in these datasets. The researchers then applied linear probes by taking the activations of the model's hidden states (the internal representations) for each entity name at various layers and training a linear regression model, or probe, to predict the actual space or time coordinates of these entities. The probe learns to decode the spatial and temporal information from the model's activations. The results showed that LLMs learn linear representations of space and time across multiple scales. These representations were robust to prompting variations and unified across different entity types, such as cities and landmarks. The researchers also identified individual "space neurons" and "time neurons" that reliably encode spatial and temporal coordinates. Above is a successful example of researchers interpreting the inner mechanism of neural networks.

Furthermore, we should have sufficient confidence that such techniques will become more established in the future. There is a famous analogy made about the NeuralPS conference: artificial intelligence today is like alchemy. In other words, both subjects are pre-scientific disciplines that gave out partially good results, but the understanding and the explanations for the phenomenon must be present. I agree with this analogy, but the implication should be positive. The world of atoms, the subject of alchemy and later chemistry, and machine knowledge are those beyond direct human perception. If we say the core issue of neural

networks today is that we cannot directly perceive their internal mechanisms, the same was the issue of alchemy five hundred years ago. However, chemists today triumph in drawing knowledge of the atomic, despite being beyond their direct perception. This triumph testifies to the possibility of perceiving worlds beyond, and we researchers of neural networks today should have confidence in completing a similar task, namely drawing information of worlds of machine knowledge.

The world of atoms and molecules, the fundamental building blocks of matter, operates at a scale far too minute for human senses to perceive directly. The atomic world is governed by quantum mechanics, a realm of probabilities and wave functions that defies classical intuitions honed by our macroscopic experiences. The interactions between electrons, protons, and neutrons occur at energies and distances that are entirely alien to the human natural understanding. In this sense, chemistry is inherently beyond humans' direct comprehension. However, chemists have developed a suite of tools and methodologies that allow them to probe, manipulate, and make sense of the atomic world. Microscopes that can visualize individual atoms, spectrometers that can decipher the energy states of electrons, or computers that can simulate the interactions between molecules.

The analogy to chemistry suggests that the inability to directly comprehend machine knowledge from neural networks does not preclude humans from understanding their functions through other methods. As chemists have built a comprehensive framework to interpret and predict chemical phenomena, AI researchers can develop interpretive methods to make neural network decision-making more transparent, like the linear probes described above. Chemistry is the testimony.

4.2 Internal Justification

For the length of this paper, I do not go on to formulate an argument for the internal justifications of neural networks but only point out how this perspective is necessary and worth future inspection.

As stated above, the current discussion on the epistemology of neural networks is centered on how humans can learn and justify the knowledge within neural networks; however, since the emergence of AI's self-awareness is a crucial topic, considering neural network as the agent itself can justifying its own knowledge is a topic with equivalent importance. That is, to consider the internal justification of neural networks is first to assume that the neural networks may have self-awareness, and then how can it justify its own knowledge? There have been many prior analyses on the importance of discussing the consciousness of machines (i.e. Chalmers, 1995).

As stated in the above section, the key to Internalist Justification is the agent's access to some mental state that generates the knowledge, and it happens that a very recent paper discovered the existence of such a mental state in "Large Language Models: Large Language Models Represent Space and Time" (Gurnee & Tegmark, 2024).

The book *Thinking, Fast and Slow* (Kahneman, 2017) by Daniel Kahneman offers a profound insight into the human mind, revealing the dual-process system that governs our thoughts and decisions. The book educates us on the interplay between two systems: System 1, which is fast, intuitive, and often unconscious, and System 2, which is slow, deliberate, and logical. Kahneman demonstrates how these systems shape our judgments and choices, often leading to biases and cognitive illusions. Such two systems of mind are philosophical faculties and can serve as the basis of internalist justification.

The paper "World Models" (Ha & Schmidhuber, 2015) integrates principles from Daniel Kahneman's book "Thinking, Fast and Slow" by leveraging the dual-process theory of cognitive function to enhance the development of artificial intelligence systems. It draws on the book's concepts of System 1 (fast, intuitive thinking) and System 2 (slow, deliberate reasoning) to design AI architectures that can efficiently

process and learn from complex environments. By mimicking System 1 with a Variational Autoencoder (VAE) for intuitive perception and System 2 with a recurrent neural network (RNN) controller for decision-making, the authors of “World Models” create a framework that allows AI to balance rapid, heuristic-driven responses with more contemplative, strategic actions, thereby enabling the AI to navigate environments with a nuanced understanding reminiscent of human cognitive processes.

Later in the paper, "Language Models Represent Space and Time", researchers discovered the existence of emergent World Models within Large Language Models. Although LLMs are not designed to have world model structures, they have world model structures that emerge within themselves after training. This signals that the LLMs, a neural network structure, have a faculty of mind, which is the key factor to internalist justification. Therefore, how AI agents justify their own knowledge must be a topic of discussion.

5 Conclusion

In this paper, I systematically analyzed the knowledge generation in neural networks, from the genesis of knowledge to characterizing knowledge and determining probable justification. The entire analysis complements a few questions that need to be addressed in previous analyses, pointed out in the introduction, and the conclusion differs from the prominent view that machine knowledge is risky.

In the first part, I demonstrated how the mathematical nature of truth is necessary for generating knowledge since neural networks are, by default, mathematical. However, whether or not truth is mathematical still awaits future philosophical discussion. There are, however, numerous famous instances that demonstrate the power of mathematics in modeling the world. Importantly, the trait as universal approximators allowed neural networks to learn mathematical truths. This explains the genesis of neural network knowledge. This is the necessary premise of neural network knowledge generation. If researchers face a problem where mathematical representation does not exist, it is not likely that neural networks will function well.

To characterize machine knowledge, I provide that neural networks bypass Wigner's layers of science and the following limitations, leading to another distinction between human knowledge and machine knowledge, in addition to those suggested by Bai (2021). This distinction further testifies to machine knowledge’s fundamental difference from human knowledge. I deem this distinction influential in how we treat neural networks and utilize their power, as it leads to the conclusion that Reliabilism is the only possible form of epistemic justification.

Reliabilism represents the epistemic justification where people trust the process of knowledge generation when a chain of logic cannot produce the result of knowledge. While intuitively, Reliabilism is not acceptable as a method of justification in critical fields such as medicine and health, I see the potential for future resolution in this issue because recent successes in techniques such as Linear Probes are helping with our understanding of the internal mechanisms in the now believed "inscrutable" black-boxes. This provides an optimistic view regarding our understanding and control of neural networks, as opposed to the now popular skeptical and critical arguments. However, this paper needs to address how people should treat neural networks before our understanding of techniques and relevant theories is complete. Concerns about safety and explainability may still occur when we use neural networks. Thus, I hereby advocate for a research focus shift from application to probing techniques, developing more methods like the Linear Probe.

In the ending section, this paper considers the importance of analyzing neural network knowledge from

the perspective that neural networks are intelligent agents since the emergent world models are a symptom of the trend, marking an emergent faculty of mind in neural networks: the ability to understand.

In summary, this paper demonstrates an optimistic view towards knowledge of neural networks. It transcends the limits of layers of science——inevitably approximations and forever pursuit of the next layer——and, as this paper predicts, its learning process and inner mechanism will be understood in the future.

References

5. Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Mohamed, N. A., & Arshad, H. (2018). State-of-the-art in artificial neural network applications: A survey. *Heliyon*, *4*(11). https://doi.org/10.1016/j.heliyon.2018.e00938
6. Bai, H. (2022). The Epistemology of Machine Learning. *Filosofija. Sociologija*, *33*(1), 40-48.
7. Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, *2*(3), 200-219.
8. Comte, A. (1876). Course of positive philosophy. *London: George Bell and Sons*.
9. Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, *2*(4), 303-314. https://doi.org/10.1007/BF02551274
10. Dirac, P. A. M. (1977). The Relativistic Electron Wave Equation. *Europhysics News*, *8*(10), 1-4.
11. DiSalle, R. (2013). The transcendental method from Newton to Kant. *Studies in History and Philosophy of Science Part A*, *44*(3), 448-456. https://doi.org/https://doi.org/10.1016/j.shpsa.2012.10.006
12. Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. *arXiv [stat.ML]*. http://arxiv.org/abs/1702.08608
13. Ernest, P. (1998). *Social constructivism as a philosophy of mathematics*. State University of New York Press.
14. Frické, M. (2015). Big data and its epistemology. *Journal of the Association for Information Science and Technology*, *66*(4), 651-661. https://doi.org/https://doi.org/10.1002/asi.23212
15. Gurnee, W., & Tegmark, M. (2024). Language Models Represent Space and Time. *arXiv [cs.LG]*. http://arxiv.org/abs/2310.02207
16. Ha, D., & Schmidhuber, J. (2018). World Models. https://doi.org/10.5281/ZENODO.1207631
17. Hammoudeh, A., Tedmori, S., & Obeid, N. (2021). A Reflection on Learning from Data: Epistemology Issues and Limitations. *arXiv [cs.LG]*. http://arxiv.org/abs/2107.13270
18. Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359-366. https://doi.org/https://doi.org/10.1016/0893-6080(89)90020-8
19. Ichikawa, J. J., & Steup, M. (2024). The Analysis of Knowledge. https://plato.stanford.edu/archives/fall2024/entries/knowledge-analysis/
20. Kahneman, D. (2017). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.
21. Makridakis, S. (2017). The forthcoming Artificial Intelligence (AI) revolution: Its impact on society and firms. *Futures*, *90*, 46-60. https://doi.org/https://doi.org/10.1016/j.futures.2017.03.006
22. Neugebauer, O. E. (1983). Exact Science in Antiquity. In O. Neugebauer (Ed.), *Astronomy and History Selected Essays* (pp. 23-31). Springer New York. https://doi.org/10.1007/978-1-4612-5559-8_3

23. Pappas, G. (2023). Internalist vs. Externalist Conceptions of Epistemic Justification. https://plato.stanford.edu/archives/spr2023/entries/justep-intext/
24. Psillos, S. (2007). *Philosophy of science AZ*. Edinburgh University Press.
25. Rescher, O. H. N. (1959). On the Epistemology of the Inexact Sciences. *Management Science, 6*(1), 25-52. https://doi.org/https://doi.org/10.1287/mnsc.6.1.25
26. Rose, N. J., & De Pillis, J. (1988). *Mathematical maxims and minims*. Rome Press.
27. Smart, A., James, L., Hutchinson, B., Wu, S., & Vallor, S. (2020). *Why Reliabilism Is not Enough: Epistemic and Moral Justification in Machine Learning* Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, USA. https://doi.org/10.1145/3375627.3375866
28. Stang, N. F. (2024). Kant's Transcendental Idealism. https://plato.stanford.edu/archives/spr2024/entries/kant-transcendental-idealism/
29. Sutskever, I., Vinyals, O., & Le, Q. V. (2014, 2014). *Sequence to Sequence Learning with Neural Networks* Advances in Neural Information Processing Systems, https://proceedings.neurips.cc/paper_files/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf
30. Ullian, W. V. O. Q. J. S. (1978). The Web of Belief. *Random House New York, 2*.
31. Wheller, G. (2016). Machine Epistemology and Big Data. In L. M. a. A. Rosenberg (Ed.), *The Routledge Companion to Philosophy of Social Science*. Taylor & Francis Group. https://doi.org/https://doi.org/10.4324/9781315410098
32. Wigner, E. P. (1990). The Unreasonable Effectiveness of Mathematics in the Natural Sciences. In *Mathematics and Science* (pp. 291-306). WORLD SCIENTIFIC. https://doi.org/doi:10.1142/9789814503488_0018
33. 10.1142/9789814503488_0018
34. Wigner, E. P. (1995). The Limits of Science. In J. Mehra (Ed.), *Philosophical Reflections and Syntheses* (pp. 523-533). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-78374-6_40
35. Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H.,…Dean, J. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv [cs.CL]*. http://arxiv.org/abs/1609.08144
36. Yampolskiy, R. V., & Spellchecker, M. S. (2016). Artificial Intelligence Safety and Cybersecurity: a Timeline of AI Failures. *arXiv [cs.AI]*. http://arxiv.org/abs/1610.07997