The Replicator Theory of Knowledge:

A Restatement and Extension

By

Almos C. Molnar

Submitted to the Department of Philosophy, Central European University

In partial fulfilment of the requirements for the degree of Master of Arts

Supervisor: Katalin Farkas

Budapest, Hungary

2021

Abstract

This thesis makes a case for a new theoretical framework with which to understand knowledge that is drawn on the synergies between Popperian epistemology, neo-Darwinian evolutionary theory, information theory and computation theory. This view has its origins in the writings of David Deutsch, and it is dubbed by the author as the “replicator theory of knowledge.” According to it, knowledge is neither justified, nor true, and it has little to do with belief, unlike how it is portrayed in the traditional philosophical conception. Instead, knowledge is claimed to be an abstract replicator: a type of information, which, once embodied in a suitable physical system, tends to remain so and casually contribute to its own copying, while most variants of it do not. After exploring the academic context from which this view emerges, the thesis outlines the theory in detail and argues for its explanatory benefits. Then, among the potential sources of contest to the view, the success of Bayesianism in the cognitive sciences is considered and its implications for the theory. Subsequently, a compromise position is developed, which can facilitate a role for Bayesianism within an extended replicator theory of knowledge.

Table of Contents

1. Introduction

One of the most persistent problems of philosophy is providing a satisfactory characterization of knowledge and of how we can attain it. In the 20th century literature, the traditional starting point is to conceive it as justified true belief (JTB) and work from there, even though almost nobody finds this "classical picture" persuasive (Ichikawa, 2017). Most modern accounts take one of two prevailing routes in order to improve the situation. The first is to complement the three conditions with a fourth one. These can be dubbed as the JTB+X approaches. Candidates for this fourth ingredient (X) are inter alia "no-false-lemmas" (Armstrong, 1973) and "no defeaters" (Levy, 1977). The second approach is to ameliorate the notion of justification or to replace it with another process that is purported to "reliably" (Dretske, 1985) or "causally" (Goldman, 1967) facilitate true beliefs. However, neither of these routes is without its issues. In particular, none of them manages to successfully free itself from the plague of the traditional JTB account: Gettier-style counterexamples.

In his landmark paper of 1963, Edmund Gettier outlined two cases that showed that there are examples of belief, which are both true and justified, and yet are intuitively unqualified for regarding them as knowledge. Since then, Linda Zagzebksi (1994) has deconstructed the strategy employed by Gettier in devising his counterexamples into two simple steps: First, start with a case where somebody has a justified false belief that also meets condition or process X. Then, alter the case so that the belief is true purely by luck. Through this recipe, Zagzebksi has shown that neither the JTB+X accounts, nor the justification-improving or -replacing approaches are immune from Gettier cases as long as they commit to a fallibilistic epistemology, which she notes is the "almost universal assumption" (p. 10). The cost of fallibilism is that no matter how one tweaks the characterization of traditional knowledge-forming methodology, there is always room left for the process to err, thereby producing a justified false belief.

In turn, these developments have by and large led to the fragmentation of the 21st century literature into a wide variety of different approaches. Some have concluded that the

lack of progress in the field demonstrates that knowledge is unanalyzable in terms of its components, and that we should instead, treat it as a conceptual primitive. This is termed the "knowledge-first" approach" (Williamson, 2000; Carter et al. 2017). Others have developed theories that allow the conditions for knowledge to change according to contextual (DeRose, 2009) or pragmatic factors (Stanley, 2005). It has been claimed that this can account for some the "shifting intuitions" associated with the Gettier cases, but this remains controversial (Ichikawa, 2017). Then, there are those who would dispense with the concept all together, portraying it as a relic of folk psychology that long "outlived its usefulness" (Papineau, 2019).

In my thesis, I will make a case for a contrary view. I will argue that knowledge is (1) an important and illuminative concept that we cannot do without, its (2) conditions are not circumstantial, it is (3) analyzable by its components, and (4) none of the three components of the traditional JTB picture are necessary or even appropriate for it. In particular, I will show this by advocating a recently advanced conception of knowledge by the physicist David Deutsch (1997, 2011) that I dub the replicator theory of knowledge (RTK).

In terms of organization, this paper proceeds as follows. After outlining the relevant academic background in section 2, I will restate the view in more explicit terms than how it currently exists in section 3. I will then move onto discussing what I deem to be the main source of problem for the theory among those that remain unaddressed, namely the success of Bayesianism in the cognitive sciences (section 4). After surveying existing responses to the dilemma in section 5.1, I will offer my own compromise position in section 5.2, which can be seen as an extension of the original theory. Lastly, I will conclude with some departing thoughts in section 6.

2. Academic Background

In order to bring out the explanatory benefits of Deutsch's replicator theory of knowledge, it is important to outline the relevant academic context, before moving onto discussing the theory itself in the next section. Specifically, four explanatory frameworks from three apparently

distinct areas of scholarly research will be reviewed here to lay the necessary groundwork. In each case, the emphasis will be on what innovative role a particular theory played in its respective problem domain. These are:

1) in epistemology, the critical rationalism of Karl Popper

2) in evolutionary biology, the gene-centered view of Neo-Darwinism, as articulated by Richard Dawkins, and its implication of "Universal Darwinism"

3) in mathematics, the information theory of Claude Shannon and its Inverse Relationship Principle, as well as the theory of computation by Alan Turing, and the physical Church-Turing Thesis

The replicator theory of knowledge emerges from the deep synergies found between these different explanatory models and principles. As such, its overall appeal stems in large part from the individual strength of each vis-à-vis their rivals in the respective academic domain.

2.1. Epistemology and Critical Rationalism

For most of the history of philosophy of science, it was maintained that science relied on the method of induction (Okasha, 2002). The origin of this thought can be traced back to the beginning of the Scientific Revolution, most notably, to Bacon's *Novum Organum* (1620). According to this view, science creates knowledge by way of making inferences from particular observations and experiments in order to form generalizations. The idea that knowledge can only, or at least, primarily, be arrived at by this process is the doctrine of empiricism, which is chiefly associated with a number of British philosophers from the 17th century onwards (Duignan, 2009). For instance, Locke (1689) claims that the human mind is a blank slate (*tabula rasa*) on which sensory information is written to accumulate knowledge. Inductivism,

which forms the core of empiricism, has also been convincing to many scientists. Newton (1713: 943) famously remarks “I frame no hypotheses” (“hypotheses non fingo”), emphasizing that a scientist should not theorize before gathering data through observation and experimentation. The philosophical movement culminated in the early 20th century with its “logical empiricism” variant (also dubbed logical positivism), foremost associated with the Vienna Circle (Creath, 2017). This holds the view that sensory data and experimental verification is not just the only way to attain knowledge, but the only way to establish anything meaningful at all (ibid).

Despite its long-standing philosophical status, the bane of empiricism has been the realization that there lies a grievous problem at the heart of induction. The most influential articulation of this concern comes from within the tradition- from Hume, best put in his *An Enquiry Concerning Human Understanding* (1748). In essence, the problem is how can we generalize universal statements from a finite set of particular observations (Okasha, 2002). Can we infer that the sun will rise tomorrow from the fact that we have repeatedly seen it do so before? The principle of induction can be summarized by such mottos as the “unseen resembles the seen” or the “future will resemble the past”, for it states that if one has the same experiences recurringly under the same circumstances, then one can “inductively infer” or “generalize” that pattern to be a “uniformity in nature” (Henderson, 2018). What justifies the belief that the sun will rise tomorrow? The fact that it has always done so in the past, says the empiricist, but that is just restating the same inductive premise (Popper, 1959). Repetition cannot help since no amount of observation of X can reach a logical conclusion about anything other than those observed instances of X, such as that, therefore, all Xs are so-and-so. Several empiricists have tried to solve the problem by establishing a set of postulates, which would justify inductive inferences, for instance, Russell (1948), but few agree that they are satisfactory. The frenzy in this regard is fueled by the notion that if nothing justifies induction, empirical knowledge is seemingly impossible. As Russell puts it, if the problem of induction cannot be solved then “there is no intellectual difference between sanity and insanity” (1946:

699). Hume himself entertained epistemic skepticism as a result of realizing the severity of the issue.

Empiricists pursue justification via the method of induction. As such, Gettier, in his paper, inter alia attributes the "traditional view" of knowledge as JTB to the empiricists A. J. Ayer and Roderick Chisholm (1963). However, the preoccupation with justificationism is present in all other pre-contemporary epistemological traditions. In fact, it can be traced back all the way to Plato, who remarks in his *Theaetetus* that true opinion is insufficient for knowledge (Paley, 2012). Some argue that the traditional view of knowledge was not the JTB conception, but "classical infallibilism", originating in Hellenistic epistemology, and carried on by Descartes to the continental rationalist tradition (see Dutant, 2015). Infallibilism is an even stronger epistemic position than justificationism, for it converts the quest for knowledge into a quest for either authority or certainty, and thus it fuels skepticism about the possibility of knowledge even more so than the problem of induction. While it is debatable, which rationalists were fallibilistic justificationists, and which were infallibilists, the tradition as a whole comes no closer in articulating a plausible account of knowledge-creation that is free of either of these commitments (Popper, 1963). Kant, who notably attempts to synthesize rationalism and empiricism, also ends up embracing a conception of knowledge as justified true belief (1781: 822).

Karl Popper is deservedly regarded as one of the greatest philosophers of science. His epistemology is dubbed "critical rationalism", and its primary innovation is the recognition that objective knowledge, while possible, is not dependent on either justification or the concept of truth in the traditional sense (Miller, 1994). Although he is not the first to hold a fallibilistic epistemology, his is the first non-justificationist one. While Popper agrees with Hume that the problem of induction is real, he does not see it as a threat to science, for he argues that in contradistinction to the classical conception, science does not rely on an inductive methodology (Popper, 1959). According to him, observation and experimentation are indeed essential to science, but their role is different than what is supposed by the empiricists. We do

not start with “pure empirical data” and infer generalizations from it, for there is no such thing. Instead, all observations are inherently “theory-impregnated”, that is, laden with a set of background assumptions (Magee, 1973). Without these, we wouldn’t be able to interpret the data. Consequently, we start with theories instead, and use experiments to test their relative merit. If a theory is rebutted by a set of experiments, then, depending on the severity of the refutation, it is either discarded, or modified according to some novel conjectures about the experimental results, upon which the new variation is subjected to further tests, and so on. Knowledge-creation is thus an alternating process of theory-formation and empirical testing, but theories come first. At the beginning of this process, far from being a “blank slate”, the human mind is already loaded with a set of inborn theories (expectations and rules of thumb), which are then repeatedly modified by life experience (Popper, 1994) These innate conjectures are bootstrapped into our brains by the instructions encoded in DNA. As a result of ontogenic development, some of these theories – as updated by the constant stream of experience – will become gradually ever more explicit; others are learned from cultural processes through language use. (See the last section for more on this).

What about the other end of this alternating process of theory-formation and empirical testing? When can a theory be finally regarded as verified, or at least, as justified as being true? What kind of observation or tests does it need, and how many? Any answer other than “no kind” and “no amount” leads us back to the problem of induction for empirical verification, as sought by the logical empiricists, is impossible. Consider what empirical test could verify the theory that all swans are white. No amount of observing white swans can prove this, yet a single observation of a black swan will falsify the theory. This shows, as Popper (1959) points out, that there is a fundamental asymmetry between verification, which is unattainable, and falsification, which is attainable. Therefore, the purpose of empirical testing is not to confirm a theory, but to refute it. Popper sometimes talked of a theory being “corroborated”, but that should be merely understood as a comparative property; a theory can only be corroborated vis-à-vis those rival theories that were actually entertained as serious candidates and yet were

refuted by experimentation (Deutsch, 1997). When a theory survives all the tests scientists can think of, it doesn't achieve any special status, it is just simply not disregarded for the time being. While it remains uncontested by a new theory that is claimed to be an improvement – upon which further experiments need to be devised to choose between them – it is rational for us to rely on it as a guide for practical action (ibid). All the while knowing, that this state of affairs is merely provisional, for all theories are expected to contain some errors. As all conjectures are fallible, the aim is to find their limits and to correct for them, but at no point the procedure is "done". It is an open-ended process.

This has implications for the concepts of truth and knowledge. According to Popperian epistemology, the problem of justificationism is that it conceives of truth and knowledge as binary categories. This leads to the mistaken question of asking what processes guarantees or at least makes it probable that we arrive at the truth (which is deemed necessary for knowledge). But a theory is not either true or not, but rather it has a certain degree of truth-content. Popper terms this verisimilitude (or "truthlikeness") to emphasize that truth, and hence knowledge, are graded properties (1963). All theories regardless of how successful they are or how resistant they hitherto showed themselves to falsification can be expected to contain falsities. One of Popper's often used illustrations of this is how Einstein's theory of general relativity superseded Newtonian mechanics (see Popper, 1994). There was hardly a more experimentally substantiated theory in physics than Newton's laws of motion until Einstein's innovation. Does this fact justify the theory as being true? If we look at what prediction the theory makes for certain celestial phenomena, then it is clearly refuted. Is it false then? If we look at the tremendous amounts of architectural achievements that was made possible only by its discovery, the theory is clearly vindicated. Popperian epistemology solves this dilemma: Newtonian mechanics has a huge scope of verisimilitude, but not as much as Einstein's theory of general relativity, which in turn, must also contain falsities of its own, but to a lesser extent. Neither theory is verified, nor justified, they are both tentative conjectures, as all knowledge is. In this view, there is no difference between "just-a-theory" and "knowledge". The only thing that

individuates knowledge is its explanatory power- how much phenomena it can account for with what level of accuracy. One day (if we work at it), Einstein's theory of general relativity will be superseded by another theory that has an even higher verisimilitude.

Note that from the point of view of how much explanatory reach a theory has, it is irrelevant how the theory come about; whether it was through the collection of meticulous data about the appearance of peas, through staring into the fire, or through an apple falling on one's head (Popper, 1959). Explanatory theories are the result of creative guesswork. They originate from "under the hood" in human brains. Constrained by empirical observation, not created by it. Popper is the first to realize this, thus providing an analysis of knowledge and knowledge-creation without reliance on justification or ungraded truth. His conception of it as an open-ended process of varying theories and their subsequent selection based on their differential success is deeply resonant with another immensely successful explanatory theory, which is the subject of the next section.

2.2. Evolutionary Theory and Universal Darwinism

Charles Darwin did not discover evolution. He discovered the mechanism of evolution, namely, natural selection. Scholars have already speculated that the diversity of life might be the result of gradual change in organisms long before him, including Darwin's grandfather, Erasmus Darwin, but that was not the hard piece of the puzzle (Ayala, 2006). The problem was identifying what drives this process. Before Darwin, the primary candidate for a solution was a theory named after the French naturalist, Jean-Baptiste Lamarck. According to Lamarckism, experiences acquired by an organism during its lifetime will be inherited by its offspring (Bowler, 2003). For example, succeeding generations of giraffes gradually acquired longer and longer necks because their forebearers constantly stretched them during their lives to reach leaves at high places. At the time, the mechanism of heredity was thought to be pangenesis- the idea, originating from Hippocrates, that each body part emitted its own particle, called gemmules, which aggregated in the gonads, and which when transferred to the offspring provided the

particles from which the offspring's respective body parts could grow out of (Ray, 2017). Since a body part altered by life experiences would produce altered gemmules, the theory was thought to provide credence to the Lamarckian concept of inheriting acquired characteristics. In fact, the view was endorsed by Darwin as well.

Yet Darwin's genius was to realize that regardless of the source of variation, once individual organisms possess different characteristics, the result will be that some of them will be more suited to survive and reproduce in their environment than their other conspecifics. Consequently, they will likely leave more copies of their heritable traits in future generations, who in turn will be able to do the same, and so on, until that trait becomes the prevailing one in the population (Bowler, 2003). Put simply, variations are "naturally selected" by the environment. Darwin coined the mechanism so to contrast it with "artificial selection"- the selective breeding of animals and plants by humans according to certain desirable traits, practiced for millennia. Since the mechanism would be automatically at force for all possible sources of variation, Darwin recognized that his mechanism was more fundamental than Lamarckism even if the predominant form of individual variation was in fact Lamarckian. This led Darwin to accept Lamarckism as a supplementary mechanism of evolution, but not as its main one (Ayala, 2006). Nevertheless, because he could not specify the source of individual variation any better than Lamarck did, his brand of evolutionary theory remained unpopular until the rediscovery of Mendel's work on biological inheritance in the early 1900s; a period that has been termed the "interphase of Darwinism" (Largent, 2009).

Classical genetics can be said to have been born, when Thomas Hunt Morgan combined Mendel's theory of inheritance with the Boveri-Sutton chromosome theory in 1915, which conclusively refuted pangenesis (Bowler, 2003). Subsequently, R. A. Fisher integrated Darwin's theory of natural selection with classical genetics in his 1930 book, *The Genetical Theory of Natural Selection*. The marriage of the two explanatory theories has been termed the "Modern Synthesis" in biology, and the resulting evolutionary theory, the Neo-Darwinian one. Refinements of the synthesis continued inter alia through the work of J. B. S. Haldane

(1924-1934) and W. D. Hamilton (1964), before G. C. Williams articulated the first explicitly gene-centered view of evolution in his *Adaptation and Natural Selection* (1966). This view was further developed and popularized by Richard Dawkins in *The Selfish Gene* (1976) and in much of his later work.

Neo-Darwinian evolutionary theory holds that the unit of natural selection is the gene. Differential survival of gene variants (called alleles) will lead to changes in their relative frequencies within the population (Dawkins, 1976). The origin of all gene variation is mutation-alterations in the DNA sequence brought on by either copying-errors in the relevant cell mechanisms or by environmental factors, such as various types of radiation. In both cases the resulting variation is "blind"; it has no foresight as to how this will affect the function of the gene or the welfare of the hosting organism (ibid). It is a commonly held, but erroneous view that evolution optimizes for the benefit of the individual. An even older mistake is that it ensures the good of the species. Neither is true. Evolution only facilitates the relative ability of gene variants to spread through the population (ibid). As such, it optimizes the gradual emergence of genetic material that is good at *replicating* itself, regardless of how that impacts its host organism (Deutsch, 2011). This is what the phrase "selfish gene" coined by Dawkins meant to emphasize. Evolution can even favor the spread of genes that are harmful to the individual's chances of survival. A textbook example is the peacock's tail, which makes it harder for the bird to evade predators (ibid). However, in practice, most genes do confer some functional advantage onto their hosts, because at least one reliable way of spreading through the population is to increase the reproductive success rate of the individual in which they reside.

Popper (1976) points out that with the benefit of hindsight, one can see that the error of Lamarckism with respect to how evolution works has the same underlying logic as the error of inductivism with respect to how knowledge is acquired. As he puts it "Darwinism stands in just the same relation to Lamarckism, as does deductivism to inductivism, selection to instruction by repetition, and critical error elimination to justification" (p. 195). Both Lamarckism and inductivism assumes that knowledge is somehow automatically present in experience and

can be mechanically derived from it. But in reality, knowledge has to be first created by some other means, independent of experience, and then, only differentially selected by it. In the case of genetic knowledge, the means is random mutation in the gene, and the differential selection is the “natural selection” of the environment. No amount of stretching one’s neck could result in stronger neck muscles (let alone a longer neck) if the knowledge on how to control muscle growth or loss depending on use and disuse was not already present in the relevant set of genes (Deutsch, 2011). In the case of scientific knowledge, it is created by intentional conjecture- by entertaining creative alternatives to existing theories, only then to be differentially selected by experiments and criticism. Due to its strong analogy to Darwinism, Popper’s theory of knowledge was the first to be termed “evolutionary epistemology” (Campbell, 1974).

What warrants talking about “genetic knowledge” is the recognition that the gene-centered view of Neo-Darwinism does not fundamentally refer to anything specifically biological. Of course, DNA is a biological substance, but its significance does not lie in the particular material it is made of, but in what it functionally enables: the storage of information of such type that cause certain environments to make copies of it (Dennett, 1995). Variation in this information for any reason, such as copying errors or information degradation, will impact the ability of the surrounding environment’s copying machinery to facilitate further replication, thus leading to the differential proliferation of the entity being copied (Dawkins, 1983). Consequently, evolution by natural selection will take place on any entity which has this property, regardless of other characteristics, such as material composition. This generalization, implied by the gene-centered view of evolution, was first asserted by Richard Dawkins (1976, 1982, 1983), and was further articulated inter alia by Daniel Dennett (1995, 2017), Susan Blackmore (2000), and David Deutsch (1997, 2011). It is dubbed “universal Darwinism”, and its chief message is that evolution by natural selection is not primarily a law of biology, but that of information theory.

2.3. Information Theory and the Inverse Relationship Principle

The oldest meaning of the term "information" can be found in the writings of Cicero and Augustine, who use the Latin word *informatio* as translation for the Greek words, *idea* (idea) and *morphe* (form), amongst others, when discussing Platonic concepts, such as the theory of forms (Adriaans, 2020). In Aristotle's doctrine of four causes, the "formal cause" of change refers to the essential pattern or shape of something that is responsible for that change. He gives the example of the numerical ratio 2:1 as being the formal cause of the musical octave, which, as Adriaans notes, "illustrates the deep connection between the notion of forms and the idea that the world was governed by mathematical principles" (ibid). Thus, already in classical philosophy, the term information (or at least, its etymological forebearer) is associated with epistemology. This connotation is also evident in the case of some medieval philosophers, such as, Augustine, whose analysis of vision features the formation (*informatio*) of the mind through the sense of sight (ibid). Here, as in the works of the classical philosophers, the term "information" is used to denote the process of being molded or shaped into form (Gleick, 2011). As such, a recurring analogy, employed by both Plato and Augustine, is to link our minds or memory to a piece of wax being imprinted with a signet ring or by writing on it (Adriaans, 2020). During the era of early modern philosophy, some have continued to make similar analogies, but no significant conceptual advancement on or even explicit analysis of the term was offered. In fact, the word "information" gradually disappeared from modern scholarly discourse; a state of affairs that remained until the first half of the twentieth century (ibid).

No one is more responsible for the modern rebirth and formalization of the term than Claude Shannon. The publication of his seminal 1948 paper, "A Mathematical Theory of Communication" (MTC), which he wrote while working at Bell Telephone Laboratories, is generally considered to be the decisive event that established the field of information theory. In it, Shannon argues that the "fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point" (p. 1). He models this process the following way. The information source chooses a particular message

out of a set of possible ones. Then, the message is encoded into a transmitter in the form of a signal and transmitted across a communication channel to the receiver, a sort of inverse transmitter, where it is decoded back to the message for the benefit of the information destination. In telephony, the transmitter is an electric device, the channel is a wire, and the signal is varying electrical current. In speech, the information source is the brain of the speaker, the transmitter is her vocal system, the signal is sound pressure, and the channel is air (Shannon & Weaver, 1949).

Regardless of the communication medium though, all channels are "noisy", or in other words, subject to error. Thus, the underlying concern of MTC is to provide a way to measure the particular characteristics of the message, signal, and channel in order to reduce noise and to enhance the efficiency of the communication system (Gleick, 2011). For instance, all channels have a capacity- the amount of signal they can transmit in a given time. Also, in many formal symbolic systems, such as in natural languages, the constituent elements of the system may be used, or may follow one another, with unequal frequencies. In English, the letter "e" appears a great deal more than "z", by far the most likely letter to follow a "q" is "u", etc. Therefore, one way to increase the efficiency of the communication system is to encode those elements of the symbolic system from which the message is constructed (e.g. letters) into such corresponding signal elements (e.g. a sequence of varying currents) that take up the least amount of the channel's capacity (ibid). Put simply, it makes more sense to use a single low-voltage current to code for the letter "e" and the sequence of high-low-high-high voltages for "z" than vice versa, since "e" will have to be transmitted a lot more often.

Shannon's conception of information transmission and its associated design tricks also apply to memory, which can similarly be thought of as an information channel (Dennett, 2017). The key characteristic of such channel is not just that it transmits the inputted information, but that it stores it. Alternatively, it can be said to transmit information not from one place to another, both from one time (the past) to another time (the future). Humans, besides their biological memory, use a wide range of physical objects as memory devices: books, hard drives, tapes,

and historically, of course, wax tablets. The only fundamental physical property memory devices have in common is that they can be all put in many different long-lived states (Tegmark, 2017). If this property is satisfied, any substrate can be employed as a memory device, highlighting that information is an abstract entity. Shannon's MTC offered a way of measuring this entity based on how many units of different states its storage - or how much channel capacity its transmission - needs, inter alia coining the term "bit" in the process, and redesignating the term "information" from its classical usage as a verb to a mass noun (Dennett, 2017). More importantly, the way it measures information is independent of what that information is about (Shannon & Weaver, 1949). In fact, Shannon's formalization says very little about what constrains what makes information so useful, namely, its semantic component.

Floridi (2010) remarks that in the philosophy of information, it has become common to define information in terms of "data plus meaning" (p. 20). Data is what is measured for communication and storage purposes under MCT. Meaning is the semantic component of information left largely unaffected by it. If we say that a notebook contains information about something in the world, we mean there is a relation between the particular state of the notebook, and a particular state of the world. In this sense we are only talking about the semantics of the information. What we state in effect is that the former *represents* the latter, that is, the information captures some of the physical properties of the real-world objects or processes in the language of mathematical abstraction (Deutsch, 2011). The representation may be only weakly abstract, such as when we denote the ratio of distances between stars in the night sky in a drawing, or more strongly abstract, such as when we do the same using natural language or binary code (ibid). It is a remarkable feature of information that the same semantic content can be acquired by different information-storing systems that share no channel or representation scheme (Dennett, 2017). As such analyzing the principles the govern the semantics of information is still a hot area of current theorizing (Floridi, 2015). Most of these efforts are conducted largely independently of MTC, but for a notable exception see Dretske (1981).

Yet MTC does constrain any theorizing of semantic information a bit (pun intended), and one of the key ways in which it does is captured by the so-called “Inverse Relationship Principle” (Floridi, 2010). The principle states that there is an inverse relationship between the predictability of something (a proposition, an event, or a state of the world) and its informativeness- the less probable something is, the more informative it is once it is so. Since a “u” almost always follows after a “q” in English, receiving it has very little informational value when trying to pin down a word starting with “q”. Instead, if an improbable “a” is received after the initial “q”, it is highly informative, for it narrows down the list of possible words to only a handful of borrowed ones, like Qabalah. Interestingly, the principle was not discovered by Shannon. It was first proposed by Popper only to be formalized subsequently under MTC (ibid). As he put it in his *The Logic of Scientific Discovery*:

> Thus, it can be said that the amount of empirical information conveyed by a theory, or its empirical content, increases with its degree of falsifiability (1935 [1959: 113]).

This is because the more ways there are to falsify a theory, the more that means that parts of it or all of it could actually be otherwise. And the more ways something could be otherwise, the more the actual way it is, can be deemed as highly specific. Therefore, theories that are tightly formulated so that they make a lot of specific predictions, while proving themselves resistant to falsification at the same time, are, in all possible sense of the term, highly *informative* about the phenomena they are trying to explain. They are, in Popperian parlance, “bold conjectures” that sticked their neck out substantially and still survived (1963). Illustrating this by going back to the example of the notebook, one can say that the accuracy (truthlikeness) to which the state of the notebook (e.g. writing) represents the corresponding state of the world is dependent on to what degree the particular information state could be otherwise (informativeness vs. probability) and the degree to which when put in one of those altered

states (e.g. by changing some words), the modeling relationship is shown to break down (e.g. fails empirical test). This property of information is reflected in the famous phrase of D. C. MacKay, “information is a distinction that makes a difference” (Floridi, 2015).

This insight also provides one of the strongest arguments for why it is right to conceive of DNA as storing information. The evolutionary lingo of “biological adaptation” refers to the fact that in the case of a gene there exists almost no such tiny change to its constituent nucleotide sequence that would make it more *adept* at performing its function (Deutsch, 2011). And while there might be a relatively small set of possible changes in the series of A-C-G-Ts that would result in no practical difference, the vast majority of possible changes would leave the gene much worse off. Therefore, both the information stored in DNA and the scientific theories stored in memory devices, such as books, hard drives and human brains are distinguished from other pieces of information by the fact that they are difficult to vary in a way that would leave their ability to fulfill their function intact. As we will see, this property of hard variability is of central importance to the replicator theory of knowledge, and it follows directly from Popper and Shannon’s Inverse Relationship Principle.

2.4. The Theory of Computation and the Physical Church-Turing Thesis

The significance of the concept of information however does not just lie in the fact that it can be transmitted and stored, but that it can be transformed in systematic ways. Transforming one particular piece of information - or information state - into another one according to some set of rules or procedure is called computation (Tegmark, 2017). As a term, computation was formalized in the 1930s by several luminary mathematicians in pursuit of answering the question what functions are effectively calculable (Copeland, 2017). In mathematics, a function is a process that associates each element of a set (the domain) to a single element of another set (the codomain). The same input to the function thereby always gives the same output (ibid). Accordingly, models of computation attempt to implement this process.

The most famous one of these was developed by Alan Turing. In 1936, he created a model of an abstract machine that could manipulate symbols fed into it on the discrete "cells" of a strip of infinite tape in accordance with a table of rules. Other important formalizations conjectured independently around the same time includes Kurt Gödel's, centered on the definition of general recursive functions, and Alonzo Church's, focused on defining functions through his lambda calculus. Church and Turing proved that these three distinct formalizations of computable functions coincided: a function is lambda computable if and only if it is Turing computable, and if and only if it is general recursive (ibid). This is known as the Church-Turing Thesis (CTT), which implies that anything that can "naturally be regarded as computable" can be computed by a "Turing machine" (Turing, 1936: 135). Turing also demonstrated that there is a special type of Turing machine, a general-purpose one, which can be instructed to simulate the behavior of any other Turing machine; specific-purpose ones dedicated to computing only one set or class of functions (de Mol, 2018). In combination with CTT, this establishes the *universality* of computation- the fact that there exists a universal computing device that can compute everything that is computable (Deutsch, 1997).

At its outset, the view was to interpret CTT as a solely mathematical conjecture, but much of Turing's own work imply a belief in a stronger, physical version of the thesis. For instance, he remarks that "[a] man provided with paper, pencil, and rubber, and subject to strict discipline, is in effect a universal machine" (1948: 416). Seen in this light, his idea of a universal computing device is based on the idealized epistemic reach of humans. He also envisions the possibility of building such devices physically with the intention of having them "carry out any operations which could be done by a human computer" (1950: 444). Furthermore, in one of his most well-known paper, "Computing Machinery and Intelligence", he argues for the possibility of artificial general intelligence instantiated in physical digital computers. Building on these ideas, many scholars have come to advocate a physical interpretation (or reformulation) of CTT, ranging from modest ones such as "anything that is physically computable is Turing computable" (Németi and Dávid 2006; Beggs and Tucker 2007) to strong ones, such as

"anything that is physically possible can be simulated by some universal computing device" (Deutsch 1985, Pitowsky, 1990). For an overview, see Piccinini (2017).

In particular, Deutsch's (1995) formulation, which has been since termed the Church-Turing-Deutsch Principle (CTD), is that 'every finitely realizable physical system can be perfectly simulated by a universal model computing machine operating by finite means' (p. 3). His conviction for this comes from his work on quantum theory. He notes that classical physics, which involves the use of real numbers and, hence, is "continuous", and the universal Turing machine, which can only represent computable reals and, hence, is "discrete", cannot obey this principle. However, a unification between physics and the theory of computation is possible under quantum mechanics: universal quantum computers, as conjectured by Deutsch and others, may obey the CTD principle, provided that quantum theory can describe every physical process there is.

Regardless of the success of quantum theory however, what physical interpretations or reformulations of the CTT (such as the CTD) aim to stress is that while the theory of computation has its historical origin in pure mathematics, to study its implications entirely in is abstract is to miss the whole point of it. As Deutsch (1997) argues the "significance of universality lies in the fact that universal computers, or at least good approximations to them, can actually be built, and can be used to compute not just each other's behavior but the behavior of interesting physical and abstract entities" (p. 98). The importance of this fact cannot be understated for it is what makes human knowledge possible.

As explained in the preceding section, the concept of information shows that physical reality has a special property of self-similarity: some of its parts can abstractly resemble its other parts, such as when a theory written in an astronomy textbook correctly explains the motions of the planets. In turn, what the theory of computation shows is that - by transforming one information state into another - this self-similarity can be improved quasi-indefinitely. Some parts of physical reality can come to resemble its other parts with ever increasing fidelity. This is only a matter of processing speed and memory capacity (Deutsch, 2011). For example, such

an improvement in resemblance is attained when an alternative to the notebook's astronomic theory is written, which turns out to have an even greater explanatory power. When this phenomenon takes place, knowledge is created, and that knowledge is the result of computation being performed by a physical entity. Consequently, the existence of knowledge in physical reality is evidence that physical computers also exist in the world. Since most knowledge (that we are aware of) is created by human brains, it is correct to consider them a type of computer. In cognitive science, this view is known as the computational theory of mind, and it allows us to bridge epistemology with information and computation theory.

3. The Replicator Theory of Knowledge: A Restatement

3.1. What Is Being Claimed?

As demonstrated in the previous sections, there are deep synergies between Popperian epistemology, Neo-Darwinism, and the theories of information and computation. Having examined these converging links, I am at last in the position to outline Deutsch's theory of knowledge, which I term "the replicator theory of knowledge" (RTK).

We have seen that evolution by natural selection works on any entity that "contributes causally to its own copying", for any variation in such an entity during the copying process will impact the resulting variant's future prospects at further copying itself (Deutsch, 2011: 93). Given enough variation and their subsequent differential selection, the result is evolved entities that are much more adept at replicating themselves than their forebearers were. Such entities are called *replicators*. In addition to genes, other examples include crystal clays, prion proteins, retroviruses, computer scripts, and, as anticipated by Popper's evolutionary epistemology, certain ideas (see Popper, 1979). For this last type of replicator, the term meme (rhymes with gene) was coined by Dawkins (1976) after the Ancient Greek word for "imitate" (*mimeme*). A funny joke, for instance, is a meme: once it gets represented in a human's brain, it tends to cause that human to tell it to others or to write it down or otherwise copy it. It is important to

note that the overwhelming majority of ideas conjured up by human brains are not memes, for they quickly "die" where they stand without any coping. But almost all enduring ideas are, such as sayings, folk stories, songs, religious beliefs, rules of thumb, scientific conjectures, etc. (Blackmore, 2000). Many of these are in fact "memeplexes", collections of symbiotic memes that reciprocally increase each other's chances of replication (ibid).

John von Neumann, one of the most brilliant scientists of the 20th century and early researcher on replicator theory, argues (1948) that replicators have three elements:

1) a coded representation

2) a mechanism to copy the coded representation, and

3) a mechanism for effecting construction within the host environment of the replicator

He was primarily interested in the possibility of self-replicating robots, still largely only a theoretical possibility at our current level of engineering technology. However, his framework is useful to adopt in order to illustrate a point of great importance; that in the case of certain existing replicators, the coded representation, as well as the surrounding copying and constructing machineries, can adopt many different physical forms in the course of their replication history. This is most obvious in the case of memes. For instance, a funny joke can exist as a neural representation in someone's brain, as a text written in a newspaper, as tiny indentations on the surface of a CD and so on. In turn, its copying and constructing mechanisms can be talking and listening, reading in one language and writing its translation in another, computer scripts executing, etc. Yet despite the different physical instantiations it is right to consider it the same replicator, because it is the particular piece of information that is the underlying cause of the copying process, not its various physical carriers (Deutsch, 2011). Even though it is less evident, this is true of DNA as well. By today, the information contained in many gene variants also exist in the form of computer memory. The information stored there could once again be copied into biological form by synthesizing such alleles from their

constituent molecules in a biotechnology lab even if these respective alleles went extinct in the meantime in their original host species. What this also shows is that a replicator's coded representation (the first element in von Neumann's conception) is also not necessarily dependent on the particular replicating and constructing machineries (the second and third elements) with which it historically co-evolved in order for it to be copied. In a world, for instance, where biologists exist who are interested in sequencing and synthesizing genes, the relevant cellular mechanisms and the behavior of the hosting organism is no longer the only route for the coded representation in a gene to result in another copy of the gene; a good example of the phenomenon of how the "replicatory reach" of a replicator can be (accidentally) expanded.

This illustrates that the real replicator, not just in the case of memes, but in the case of genes too, as well as all other replicators beyond the most primitive ones, is in fact *abstract*: it is the coded representation, or in other words, the information, itself (ibid). The ultimate explanation for the particular way the replicator's information content, or coded representation, got to be the way it is, is always that variants of it were less successful in getting themselves replicated. As pointed out above, almost all possible small changes in the coded representation, such as a missed semicolon in a self-copying computer script, will make it substantially less *adept* in being able to replicate itself, while very few will make it more successful. Thus, shrinkage in the "replicatory reach" of a replicator is by far the most common fate as a result of variation.

The property of being hard-to-vary in its information content is what ties all examples of successful replicators together, whether we are talking about jokes, genes, computer viruses, or scientific explanations. This brings us finally to the following conception of knowledge:

> Knowledge is an abstract replicator: a form of information which, once embodied in a suitable physical system, tends to remain so and casually contribute to its own copying, while most variants of it do not. (Deutsch, 2011: 95)

What may be the most striking feature of this characterization for many is that it substantially broadens the scope of human ideas that count as knowledge. It does not just cover such memes as rules of thumb, pearls of wisdom or scientific theories, but jokes, songs, expressions, and so on as well. Even more dauntingly, it includes enduring ideas that most of us today recognize to be wrong, both in the factual and moral sense, like misogyny. However, this has to be so, for memes - as we have seen in the case of genes - do not optimize for conferring true beliefs about the world or any other functionality onto their hosts, but only for how to get themselves replicated. Much like with genes, this is achieved by encoding and exploiting some regularity in their host environment, which means they will also incidentally end up storing knowledge about that regularity in the environment. In the case of jokes, for example, this is about the particularities of the psychological capacity of humans known as a sense of humor as well as relevant affordances in their ontologies that feature in the content of jokes (e.g. horses and bars). In the case of scientific theories, the memes exploit the desire of humans to control the world around them, in which case, the successful ones have to encode some regularity about how the world works, with more successful ones having to encode more. This may be true in the case of rules of thumb and religious beliefs as well albeit to a much lesser extent, for they also tend to encode and exploit other regularities in their surrounding environment, such as that human memory retains propositions that rhyme more easily, or that people have a tendency to overattribute agency in the realm of things around them. And so, a less accurate rule of thumb that rhymes may preferentially proliferate at the expense of its more accurate rivals that have the bad luck of being expressed in unamusing prose. In the same vein, an explanation of the change of seasons that feature the familiar whims of various gods may be culturally preferred for generations over other religious accounts that are

nevertheless more mechanistic even if the latter ones have a higher predictive power. It is only *scientific theories* that solely "specialize" on the replicating strategy of approximating laws of nature to comparatively high degrees of accuracy in order to spread in their immediate host environment of human culture and its physical artefacts. But even in their case, they have to compete for the limited attention-span and memory capacity of human brains with a swarm of other memes evolved to embody replication strategies that press on alternative aspects of our psychological predispositions. The state of that competition, for instance, was very different in Europe during the Dark Ages than after the Scientific Revolution.

As is the case with viruses, memes can exist in all three types of symbiotic relationship with their hosts (Dennett, 1995). Many memes are mutualists, offering valuable ways of behaving, such as knowing how to make a sling or prepare a dish for example. Others are commensal (e.g. different pronunciations of a word) or parasitic (e.g. religious self-harm practices). In the case of certain memes, it may be highly controversial to establish, which of these categories they belong to. And what may have been a mutualist meme at some particular period in our species' history, may no longer be anymore, much like in the case of certain genes. For instance, our innate drive to be attracted to foods with high sugar content has been advantageous to us in the environment most of our hominid ancestors lived in where such food sources were scarce, but it has become decisively maladaptive today when that is no longer the case. Yet, even from enduring parasitic memes, scientists - by decoding the knowledge such memes use to replicate - may learn a lot about (potentially archaic or undesirable) facts of human psychology or of particular human cultures, if nothing else, which are themselves regularities in nature.

Consequently, just as Popper points out that there is no clear separating line between just-a-theory and knowledge, we can submit that there is also no such thing either in the case of one type of meme and another. Replicators always encode knowledge about something regardless how parochial that may be. Some endure a vast number of copying with great fidelity in a wide variety of environments, which is a sign that they encompass a great deal of

knowledge. Others break down in a few generations, which is the mark of how easily variable their information state was, and hence, how limited was the knowledge that they contained. Once again, non-categorical thinking should be adopted when contemplating the informational constitution of replicators: there is a continuum of information states from entirely non-replicating ones to full-blown replicators, with quasi-replicators, proto-quasi-replicators, and so on, all the way in between, with information becoming more and more invariant towards the high end of the spectrum. Our definition of knowledge reflects the virtues of such non-categorical thinking.

3.2. What Is Being Gained?

As advertised in the introduction, this theory of knowledge is radically different from the justified true belief account of knowledge and its derivatives. According to it, nothing justifies knowledge for the processes that create it can be vastly different, e.g. natural selection, an arbitrary quirk of human psychology, computer software, etc. The only measure of success of a particular piece of knowledge is how well it proliferates relative to its rivals in the environment that instantiates them. Knowledge, in this light, also departs from the connotation of ungraded truth, for it is based on the coded representations of some particular set of features of the hosting environment, the scope and accuracy of which is always a matter of degree that can be improved quasi-infinitely. Finally, this conception has little to do with belief in the traditional sense, for, knowledge cannot just, as Popper (1979) puts it, exist “without a knowing subject”, but without any associated mental states either. It is, of course, necessarily reliant on informational states, but unless one is prepared to equate the two, and talk of retroviruses as having beliefs in their RNA, we better disperse with it.

Nevertheless, one may still ask: why should anyone take this new theory of knowledge seriously? How is conceptualizing knowledge in these terms is an improvement in our explanatory understanding of the phenomenon as compared to existing accounts? Many of its virtues were already indicated. This conception of knowledge bypasses the problem of

justification, and with it, vulnerability to Gettier cases. Moreover, it does this without giving up the meaningfulness, the objectivity and the analyzability of the concept. As discussed in the previous sections, it also emerges from the converging ideas of several highly successful explanatory frameworks; a sign in itself that it is worth taken seriously. Yet, in my opinion there are two major benefits of this view in addition that are important to emphasize at further length.

The first is that it offers a naturalized and non-anthropocentric view of knowledge that is more unifying than under any other rival conceptions. Knowledge is no longer tied to some quirk of human mentality, such as forming “clear and distinct” understandings of something, or deriving perceptual information from our sense organs. Instead, the theory allows us not just to grant non-human animals with knowledge, but, under the traditional view, to non-living things as well. It would be misleading to conclude however that this notion thus forces us to treat knowledge to be a more fundamental phenomenon in nature than life itself. Rather, what significant about this implication, is that it offers us a novel criterion with which to demarcate what life is that far outdo traditional candidates, such as growth and development, homeostasis, energy processing or locomotion: life is knowledge-creation (Deutsch, 1997). Since the role of self-replicating molecules evolving into ever more complex replicators is prominently featured in the prevailing scientific theory of abiogenesis, this points to yet another converging link between the replicator theory knowledge and a successful scientific conjecture from a seemingly distinct discipline. Furthermore, as mandated by the universalities of Darwinism and computation, this must hold true for other life forms as well that may exist in the universe. In so far there are other instances of life and intelligence out there, their origin and evolution must have been and continue to be governed by the same fundamental epistemological principles that determine the conditions under which knowledge-creation is possible here on Earth. Far from being an outdated relic of folk psychology as some have come to believe, knowledge, as properly understood, occupies a central place in the cosmic scheme of things.

The second crucial reason why we should take the replicator theory of knowledge to heart is that it offers a crucial insight into how scientific progress is attained. According to this view, successful replicators are characterized by a highly invariant informational structure. Since scientific theories are memes that are specialized on replicating by encoding large-scale regularities in nature that humans find to their benefit to discover, successful scientific theories will be comprised of explanations that are very hard to vary without affecting their ability to account for the natural phenomena they are highly adapted to account for. In this light, progress in science is dependent on finding invariant scientific explanations that have ever broader and deeper explanatory reach.

This is an important improvement on what most people believe to be Popper's influential criterion of demarcation is; namely that science is made up of falsifiable theories (Magee, 1973). While this is true, it cannot be the whole story, for as Deutsch (2011) points out, falsifiable theories have always been common. Any gambler who believes they know what the next roulette number will be has a falsifiable theory, and so does every self-styled oracle who augurs the forthcoming of a particular event. However, their theories are not scientific, because making predictions is not the purpose of science. The purpose of science is finding explanations, for appearances are not self-explanatory (ibid). Thus, having an explicit explanatory content in addition of being testable is also a necessary characteristic of genuine scientific theories. Yet, this is still not sufficient to demarcate them from non-scientific ones, for even falsifiable, explanatory theories have been abundant in the history of our species (ibid).

Consider religious myths, which have long provided explanations for just about every natural phenomenon of human interest. For example, in Nordic mythology, people accounted for the fact that the constellations do not move relative to one another in the night sky by claiming that they are fiery sparks affixed onto the skull of the slayed frost giant Ymir, in which we live (Davidson, 1964). This explanation, as all explanations, posits distinct claims about reality, such as that the sky is a dome we perceive from the inside, that the dome is the skull of enormous being, etc. that are, in principle, all testable. Why is it then that such testable,

explanatory theories fail to qualify as genuine scientific theories? The reason is that the particularities of the explanation barely connect to the particularities of the phenomenon they are trying to explain. Why the skull of a “frost giant” and not a bear? Why a skull at all rather than a similarly shaped object, like an apple? And so on. Changing any aspect of the explanation along the lines I offered would account just as well for the phenomenon. In other words, almost the whole of the explanation is easy to vary, along with the background assumption it relies on (like the existence of various gods). This means that when any of the testable predictions of such easy-to-vary explanations were to be proved wrong by observation or experiment, its advocates could effortlessly modify that part of the explanation, all the while getting almost nowhere closer to the truth. This is why the vast majority of testable explanatory theories can be rejected out of hand without any attempt to falsify them just on the grounds of being easy-to-vary explanations (Deutsch, 2011). In contrast, according to RTK, successful scientific theories are not just testable and explanatory, but hard-to-vary in their explanatory structure (i.e. they are highly adapted to their purpose). There is no easily implementable change to the laws of thermodynamics or neo-Darwinian evolutionary theory that we can resort to if they are falsified by experiment. Its advocates cannot just change a few words of it and expect it to be an advancement. Improving upon them is an exceedingly difficult task that has little to do with more observation and empirical evidence, and almost everything to do with creative conjecture.

One might detect a hint of contradiction in designating myths as being easy-to-vary. After all, the only reason we know of such memes, or rather, memeplexes, today is because they were copied enough to be preserved, which means that they must have a type of invariancy in their information content that is the mark of successful replicators. However, the point is that the invariancy in such a memeplex as the slayed frost-giant theory of the stability of constellations embodies more knowledge about what was salient in the ontology of people from a warrior-culture living in cold climates than about astronomical facts.

I would like to highlight one last point before moving onto the next section. Deutsch himself is ready to stress that this criterion of demarcation - that science is comprised of testable, *hard-to-vary* explanations – can already be found in Popper's work, albeit less explicitly, for instance, in his characterization of science as an evolutionary process and in his Inverse Relationship Principle, as discussed. To what extent Popper was aware of this may be of interest to historians of philosophy, but are not really important for our purposes here. What Popper clearly states though is that the falsifiability of a theory only matters after it has survived rational criticism, which includes inter alia examining that it is logically consistent, that it is free of unnecessary complexities, etc. (Miller, 1994). If you understand this process to include evaluating whether the theory meets Deutsch's hard-to-vary principle as well, it renders the charge of critics of Popperian epistemology baseless that falsification of a scientific theory is impossible, because the empirical evidence is always logically consistent with an infinite number of theoretical variants (see Jeffrey, 1975; Howson and Urbach, 2006). As an example, one may rightfully argue that the empirical evidence we obtained thus far is logically just as consistent with the prevailing explanation for gravity, namely the general theory of relativity, as it is with the same theory with the only minor change that the next time the Queen of England jumps, she will float away. However, this has no bearing on the issue of regarding the prevailing theory of gravity as a genuine scientific theory, while the modified one not. For the prevailing theory is hard-to-vary in its explanatory content, while the alternative offered breaks this distinctive property down by introducing an unexplained element with no functional role. As Deutsch (2011) puts it "inventing falsehoods is easy, and therefore they are easy to vary once found; discovering good explanations is hard, but the harder they are to find, the harder they are to vary once found" (p. 26). This principle of epistemology, that progress in science is dependent on finding ever broader and deeper invariant explanations, flows straight from the replicator theory knowledge.

4. The Replicator Theory of Knowledge: A Challenge

4.1. The Rerise of Justificationism

We have seen that RTK offers a broad framework with which to understand various instances of epistemic success, whether we are talking about how the knowledge stored in DNA governs the growth and behavior of a biological organism, how a self-modifying computer virus spreads through the internet, or how scientific progress is achieved by our species. However, in order for it to be a truly unifying framework, it also has to converge with - or offer a worthy alternative to - our best explanations of how most cases of ordinary learning is achieved by humans; as studied by the cognitive sciences. It has been argued that RTK is deeply grounded in the computational theory of mind. Thus, it should be adept at explaining learning in any information-processing system, whether we are talking about animal brains (including ours) or artificial intelligence (AI) software. Yet, the explanatory framework that is currently sweeping through cognitive psychology and AI research in this regard with rather impressive results, is - on the surface at least - seems quite incompatible with it. This is the framework of Bayesian inference.

The notion that the processes that drive epistemic success in science and in cognitive development should be related has occurred to many. It has been an influential idea of cognitive psychology, first proposed in the 1980s, that children are natural scientists; constantly updating their conceptual structures about the world in light of their experiences (see Carey, 1985; Gopnik, 1988; Wellman, 1990). Often called “theory theory”, the view proposes that ordinary mental representations are like scientific theories and that cognitive development is the successive revision of such theories based on the emerging evidence, much like how that is done in science (Gopnik and Wellman, 2012). Of course, this analogy leads one right back to the question of just how exactly that is done in science. Popperian epistemology offers one answer. Bayesian epistemology offers another. In the past twenty years, Bayesianism has become to dominate the field (see Gopnik and Tenenbaum, 2007; Tenenbaum et al. 2011;

Clark, 2013; Hahn, 2014). As such, “theory theory” has also been reframed by its proponents in explicitly Bayesian terms (see Gopnik and Wellman, 2012).

Named after the reverend Thomas Bayes, this method of inference consists of reallocating credibility across one’s set of beliefs according to one’s prior experiences in addition to the new evidence (Howson and Urbach, 2006). Grounded in probability theory, the framework offers a normative approach in describing what would be rational for an agent to believe or do (ibid). In spite of the frightening mathematical complexities that characterize most Bayesian computational modeling done today; it is all claimed to be based on one simple equation:

$$P(H \mid E) = \frac{P(E \mid H) \times P(H)}{P(E)}$$

Known as Bayes’ theorem, the formula expresses that the posterior probability of hypothesis H given evidence E is the product of the likelihood of evidence E given hypothesis H and the prior probability of hypothesis H divided by the marginal probability of evidence E (Donovan and Mickey, 2019). The main benefit of the theorem is brought out when there are multiple hypotheses at play. While the equation gets more and more complicated as more hypotheses are entertained (especially, the denominator), as long as one is able to attribute numerical values to each of the variables involved, Bayes theorem provides a straight-forward way in deciding, which hypothesis to believe: whichever is most probable based on the available evidence and our prior knowledge (ibid).

Assigning values to the prior probabilities (or priors for short) has been the most contentious point of the procedure, with a long debate between “objectivists”, who wished to establish interpersonally neutral ways of initializing them depending on various mathematical stipulations, and “personalists”, who argued that assigning them is a matter of intrapersonal subjectivity (McGrayne, 2011). By today, the second camp has won. For explicit calculations, usually one is asked to consider how much they are intuitively willing to bet on a particular hypothesis or outcome vis-à-vis the others and have the priors initialized accordingly (ibid). As

common sense suggests, “strong” priors, ones that are heavily biased a particular way, will result in posterior probabilities (or posteriors) that are less moved by any new evidence. By contrast, when one adopts “weak” priors, the evidence will tend to dominate one’s posteriors (Donovan and Mickey, 2019). More importantly, introducing such subjectivity into the process allows one to account for individual differences in belief when confronted with the exact same epistemic situation based on the individuals’ novel set of life experiences (Gopnik and Wellman, 2012).

Regardless of such details though, the main idea behind Bayesian epistemology is that instead of seeking either conclusive verification or falsification, one should pursue probabilistic confirmation (Howson and Urbach, 2006). The primary motivation for taking this route was already hinted at above: the set of available empirical evidence always “underdetermines” the theory (Jeffrey, 1975). This is in fact the problem of induction in disguise (Longino, 2016), even though some philosophers would like to think the two issue as separate (see Duhem, 1954). Almost every epistemological tradition, Bayesianism included, agree that there cannot be a principle of induction that can logically entail moving beyond the observed data as scientific theories routinely do (Howson and Urbach, 2006). Yet, with the exception of falsificationism, most schools take this to mean that the real problem of induction is not to provide a strictly speaking valid principle of induction, which they grant to be impossible, but rather, having granted that it is logically invalid, how can one nevertheless fill in the gap and come up with procedures that result in at least “reliable” or “probable” inductive inferences (Deutsch, 1997). In other words, they pursue justification. Bayesians found their own preferred way of doing that relying on Bayes theorem (see Griffiths and Tenenbaum, 2009).

4.2. The Empirical Evidence for Bayesian Inference

The reason why the appeal of Bayesianism is so widespread in cognitive science and in AI research today is because the “problem of underdetermination”, as known in philosophy, seems especially hard-hitting (Gopnik and Wellman, 2012). In fact, the argument has been

known in these fields under various other names: linguists talk about the “poverty of the stimulus”, and AI researchers about problem of “sparse data” (ibid). Whether one is talking about object recognition, language acquisition or attributing causation, the available evidence is almost always compatible with multiple generative sources (Gopnik and Tenenbaum, 2017). Yet in the case of all of these cognitive tasks and many more, children learn to make powerful generalizations and abstractions given just a few data points (ibid). For instance, a two-year-old upon acquiring the word “horse” into her vocabulary can use it in novel situations with great accuracy (albeit not perfectly), even if she had only heard a handful of sentences that contained it (Tenenbaum et al. 2011). How does she know the boundaries of the category from just the few cases observed?

Since the Bayesian inference framework offers a clear way to mathematically formalize such cognitive tasks, it has been prevalently used to model how learning might take place in regards to them with seemingly very promising results (Hahn, 2014). By today, an impressive number of behavioral experiments show that the particularities with which the children perform on many of these tasks closely matches how an “ideal Bayesian learner” would perform (Clark, 2013). Such convergence between the output from computational models and the psychophysical data has illuminated the workings of virtually all of our major cognitive capacities, including perception, language use, memory, sensorimotor system, and various aspects of higher-level cognition (ibid).

In addition, several neuroscientific findings seem to support the view that the brain has functionally evolved to instantiate something akin to the Bayesian algorithm. For example, this explains why there are more outbound neural pathways for sensory signals than inbound ones in the visual cortex- the brain continuously feeds forward its probabilistically generated expectations and evaluates its prediction errors by matching it to the incoming signal (Dennett, 2017). Also, it provides an explanation for the spontaneous activity in the cortex, which has been observed to be approximating the average evoked response of the respective cortical areas, much like how a series of prior-posterior updates is supposed to converge on the true

probability distribution of the analysandum in question (Berkes et al. 2011). Yet as many of the underlying hardware details are still missing, Bayesianism today is largely seen as best utilized in providing - in the terms of Marr's three levels of explanations - a "computational level" explanation, rather than an "implementation level" one (Tenenbaum et al. 2011). Nevertheless, that has not deterred many to conclude that the brain must instantiate Bayesian hierarchical predictive coding networks with which it can generate and revise conceptual structures on the fly (Clark, 2013).

Finally, recent successes in AI research also lends credence to Bayesianism. At present day, the most cutting-edge AIs are all based on "deep learning", a set of machine learning techniques that - inspired by the brain's architecture - work by instantiating "artificial neural networks" (Goodfellow et al. 2016). Many of these deep learning methods are designed to approximate Bayesian inference, as the problems they aim to tackle are thought to be the same: extrapolating generalizations from large chunks of fed data (Elton, 2021). Such Bayesian deep learning models have been shown to match, and often exceed, human-level performance on a variety of cognitive tasks, such as natural image classification, driving, medical diagnosis, and playing games, like chess, go, jeopardy, and various video games (Tegmark, 2017).

Thus, it would seem that the case for Bayesian inference being involved in cognition is strong. But could it be that this framework by itself is enough to account for all aspects of what the human brain does when it creates knowledge? Not just in the cases of ordinary learning, but when it discovers a new and more powerful scientific theory as well? For many Bayesian epistemologists, the answer was already theorized to be yes, even before we had access to the aforementioned empirical data (see Jeffrey, 1975, Rosenkrantz, 1977). By today, several Bayesian researchers seem to have joined in the affirmative. For instance, Tenenbaum and colleagues (2011) argue in their suggestively titled piece *How to Grow a Mind* that hierarchical Bayesian modeling can "address some of the deepest questions about the nature and origins human thought" (p. 1). In turn, Friston et al. (2017) are keen to characterize the "aha" or

"eureka" moments associated with scientific discovery as the result of Bayesian computations generating new conceptual structures. DeepMind's Marcus Hutter (2000) claim that *all* possible cognitive problems that an AI might face can be solved by implementing Solomonoff's theory of induction, which is generally viewed as the computational formalization of pure Bayesianism. The sentiment seems to be shared by those at OpenAI; the company is set on achieving artificial general intelligence (AGI) by trying to scale up existing deep learning techniques (Goertzel, 2020).

If such Bayesians are right, then this would pose a considerable problem for Deutsch's replicator theory of knowledge, which, as discussed above, is heavily grounded in Popperian epistemology. The next section explores how one might respond to the Bayesian challenge of RTK, including Deutsch's own position. Since his stance is that of firm rejection, which I do not find entirely defensible, I will also make a case for my own solution that is of a more conciliatory nature.

5. The Replicator Theory of Knowledge: An Extension

5.1. Possible Replies to the Bayesian Challenge

In response to the contest set out above, one may legitimately decide to side with those Bayesians who claim that their framework can account for all cases of epistemic success. Accordingly, one may argue that RTK is only as viable as it conforms with Bayesianism- it is this epistemological tradition, not the Popperian one, on which it should be based on. On this view, to the extent that Popperian epistemology get things right is only because it agrees on those things with Bayesianism.

In fact, several scholars highlight ways in which parts of the former might be a valuable heuristic of the latter. For instance, the AI researcher Eliezer Yudkowsky (2006) argues that Popper's insight that there is an asymmetry between falsification and confirmation can be derived from Bayes' rule. If the probability of some evidence E given hypothesis H equals to

one (i.e. P(E|H = 1) then observing ~E falsifies H. In contrast, observing E does not confirm H for E can be consistent with some other hypothesis, say, hypothesis Z, that makes the same prediction: P(E|Z = 1). In order for observing E to confirm H, we would have to know that P(E|~H) = 0, but that is impossible, because we can't enlist every other possible hypothesis (ibid). A similar point is made by the Bayesian epistemologist Jon Dorling in his debate with the critical rationalist David Miller (Dorling and Miller, 1981). In the same vein, several other scholars have formalized other dictums of Popperian epistemology in Bayesian language, such as "subject hypotheses to severe tests" (Milne, 1995), or have emphasized other elements of similarity, such as the way they apply the "simplicity criterion", often known as Occam's Razor (Rosenkrantz, 1977). Moreover, Crupi and Tentori (2014) point out that the Inverse Relationship principle is also adhered to by Bayesian confirmation theory.

However, this is selling Popperian epistemology short. Piecemeal similarities aside, the virtue of critical rationalism goes well beyond the fact that aspects of it can be interpreted as special cases of Bayesian epistemology. As we have seen, one of its chief merits is that by taking an anti-justificationist stance in regard to how knowledge is created by human brains, it converges with the explanation of how the same is done in the genetic material by natural selection, which, as discussed, is Darwinian, not Lamarckian. Both Neo-Darwinism and Popperian epistemology hold the view that "instruction from without" is impossible in stark contrast to pure Bayesianism (Popper, 1979). Moreover, even advocates of the Bayesian framework concede that one of the lingering questions in their line of research is how the process of Bayesian inference gets started (Tenenbaum et al. 2011; Gopnik and Wellman, 2012). Some knowledge must already be there at the beginning of the procedure to inform the priors and what counts initially as evidence in order for the whole thing to be able to take off. This knowledge has to come from somewhere, either as a result of genetic evolution or cultural evolution or both, which is a question only addressed by critical rationalism among the two. This means that Bayesian epistemology cannot by itself account for all types of knowledge-creation, which is the purpose of RTK. It can only be part of the answer, not the full answer.

Another possible position to take in response to the Bayesian challenge is that of Deutsch himself, which lies at the other extreme of the spectrum. Following Popper's line of reasoning, Deutsch argues (2012; 2019) that Bayesianism is an entirely wrongheaded approach to take in trying to understand how our brain creates knowledge. His criticism is especially targeted at the field of AI research, and its much-heralded recent breakthroughs, on which he is often asked to comment. Despite the appearances, Deutsch claims that the field is no closer as a whole in attaining AGI than it was sixty years ago, even though the universality of computation mandates its possibility (ibid). He argues that the deep learning models we have today are contaminated with preexisting knowledge that are smuggled in by the programmers, and even then, they require enormous amounts of data to train, that is entirely unlike human learning (2011). Furthermore, as pointed out also by others, these AIs still fail to extrapolate beyond the boundaries of the task. When the parameters of their job description are changed even a little, they break down in major ways (Elton, 2021). For example, Google's AI providing medical diagnosis, which outperformed doctors in laboratory settings, has completely failed in field trials with different lighting conditions and image resolutions (ibid). Similarly, DeepMind's various videogame-playing AIs beat all human players, but only with those game settings on which they were trained (ibid). Adjusting them even slightly, completely demolishes their performance. Deutsch (2012, 2019) reasons that the lack of real progress in the field shows that pursuing AGI by trying to program ways to make inductive inferences along the lines of Bayesianism will bring us no closer to attaining the goal. Instead, that would require us to know how to program "creativity", which is a qualitatively different functionality; unobtainable by merely "scaling up" existing approaches (ibid). Deutsch claims that first we have to better understand how creativity works, for which he thinks there is "nothing less than a breakthrough in philosophy is needed" (2012). Moreover, according to him, that breakthrough will have to be grounded in Popperian epistemology, for as Popper explained creativity has nothing to with making inductive inferences (ibid). In conclusion, for Deutsch, Bayesianism has no part to play in explaining any epistemic success.

This is where I intend to depart from Deutsch's views, which strike me as just as overly dismissive as pure Bayesianism is, only in the other direction. Based on the empirical evidence available today (and surveyed above), I think it is implausible to conclude that Bayesian inference has nothing to contribute in accounting for knowledge-creation. It may well be the case that both it and critical rationalism merely touches on different aspects of the same puzzle. Such a view is taken by Elton (2021), for instance, who speculates that the brain may host two different types of knowledge-generating process. One could be characterized as the "brute force inductive fitting of big data", which starts with random weights and priors and gradually refines its models based on training loss signals (p. 11). This process is grounded in Bayesian epistemology. The other process is grounded in critical rationalist epistemology and it starts with models found by "evolutionary algorithms". These are rejected "whole cloth" when falsified, but have the potential to discover hard-to-vary explanations (ibid). Additionally, according to Elton, these two processes seem to map onto Kahneman's distinction between System I and System II (see Kahneman, 2011). I commend the conciliatory nature of Elton's approach, but I do not find his "dual-process" account persuasive. Partially, because it seems to me to be a dubious interpretation of Kahneman's work, and partially, because it introduces more complications than it solves.

Nevertheless, I believe that looking for a solution that accommodates both epistemological camps to some extents is on the right track. It is time for me to make a case for one such possible compromise position.

5.2. A Case for Compatibility: Bayesian Hardware and Memetic Virtual Machines

My attempt at reconciling RTK with Bayesian inference is inspired by Dennett's account (2017) on how the human mind evolved to be. Dennett argues that the available empirical evidence makes it likely that the underlying architecture of animal brains, including ours, is in fact made up of hierarchical Bayesian networks. However, he also contends that such networks by themselves cannot be the reason for the kind of reflectiveness and creativity that marks our

species (ibid). This is because while Bayesian networks are highly adept at extracting the information from the environment that matters to the organism - thereby generating affordances in abundance -, their competence exists without them needing to represent or express the reasons they track (ibid). Having reasons, beliefs, and values, all of which are characteristics of human comprehension and top-down cognitive control, are not needed to make them work. The creation of such cognitive competences is only made possible by an entirely different kind of process, namely cultural evolution (ibid).

Cultural evolution deals in cultural information, which offers qualitatively new kinds of affordances to be taken up by brains: ones that already contain knowledge. Thus, Popperian evolutionary epistemology is vindicated after all, for memes, according to Dennett, are an intimate part of the solution. Memes are cognitive competencies that are designed by cultural evolution, and are picked up onto the brain by hierarchical Bayesian networks. Since the knowledge they bring with them is not something that would have been also available for the organism by genetic means or by mere inference of statistical regularities in the environment, they can be thought of as “software” installed onto our brains. In fact, it is more accurate to consider them as virtual machines that are being run on our Bayesian hardware, for once they are installed, one of their functionalities might be to open up the possibility for further kinds of affordances to be picked up, which may be new kinds of memes themselves. Much like how when one virtually hosts a Windows on a Mac, it allows one to download and execute software that only runs on Windows. But what if the software in question is in fact another virtual machine? This opens up the prospect of nested virtualizations, with layers of virtual machines built on one another potentially ad libitum. With regards to an example with memes, the class of memes known as “phonemes” provides a way for digitizing cultural information in the form of words. Then, the new class of memes known as “words” opens up the possibility for language to emerge with which one can convey information with displaced reference, counterfactuals, etc. This in turn creates the opportunity for complex belief systems to form, such as religious doctrines or scientific explanations, which are yet another example of a novel

class of memes. This runaway process of memes facilitating cognitive competences with which other kinds of memes can be designed and picked up is what, according to Dennett, effectively turns our brains into “minds”, with all its characteristics features, such as having reasons, beliefs and values (ibid). It also what allows for the adoption of truly novel or creative perspectives, something which cannot come by making inductive inferences alone.

Thus, we cannot explain the epistemic success of humanity without cultural evolution. If animal brains are indeed instantiating approximated hierarchical Bayesian networks as the evidence suggests they do, and if the possession of such networks were to be sufficient in itself to create large-scale explanatory knowledge, then we would see other animal species doing that. But this is not the case. Of course, there are other species that demonstrate rudiments of culture. By today, many animal behaviors that were thought to be instinctual have been discovered to be habits learned by the offspring from its parents, or by observing other members of one’s species (Dawkins, 1982). Nevertheless, these remain profoundly limited. Only homo sapiens possess cumulative culture (Dennett, 2017). The pure Bayesian picture misses what is so distinct about our species, namely that for the past couple hundred thousand years the dominant form of evolution that shaped our cognitive trajectory was cultural and not biological.

What made the difference in our case? This remains an area of active research with a lot of unanswered questions. As Richerson and Boyd (2004) note in their book dedicated to emphasizing how culture transformed human evolution that “the existence of human culture is a deep evolutionary mystery on a par with the origin of life itself” (p. 126). There are many theories as to what made the difference, for instance, bipedality. Walking on two legs allowed a greater role for tool-making as well as carrying them, while also creating more opportunities for gesturing to develop for communication purposes (ibid). Another explanation favors social intelligence or theory of mind (TOM); the ability to recognize others as having information and intentions on their own and to predict their future behavior based on these. Michael Tomasello (2014) argues for example that such perspective taking may have originated under competitive

evolutionary pressures and then have later evolved into a mechanism for cooperation among conspecifics by allowing joint attention and joint intentionality. Language is yet another candidate, preferred by Dennett (2017) amongst all the others. Most likely though, all of the above and more factors served as evolutionary thresholds that prevented other species from starting to produce culture on par with us. In the case of the earliest meme transmissions, one can easily imagine that some have required bipedality, some joint attention, and some proto-language use.

Regardless of the details how cultural evolution got started though, once it did, it provided another route by which human evolution could take place besides the genetic one. From that point onward, human behavior and human cognitive competences were shaped by the interactions between these two evolutionary processes; a view known as dual inheritance theory (Richerson and Boyd, 2004). Once culture got on its way, cultural information has become just as much part of the affordances of humans as any other previous statistical regularity offered by the natural environment and harnessed by Bayesian neural networks. In the terms of Wilfrid Sellars, they became part of our manifest image (1962). Consequently, cultural evolution has also put selective pressure on the biological architecture of human brains, favoring those model variations that were better at detecting and picking up these new kinds of affordances. This in turn put on new selective pressures on memes, and so on, in a kind of never-ending feedback loop (Dennett, 2017).

An implication of this process is that the dynamics with which cultural evolution works itself has evolved over time. It started out being profoundly Darwinian just as biological evolution is. However, it gradually got de-Darwinized as more and more kinds of cognitive competences were honed by it (ibid). The earliest pieces of cultural information born around the dawn of human culture could be barely considered replicators. Much like the earliest replicating chemicals, they must have been astonishingly bad at affecting a copy of themselves. Our ancestors hosted such proto-memes with no more comprehension than what they had for the instincts they were equipped with by genetic means. However, culture-borne

pieces of knowledge were different in one crucial way even at the early days of culture: they could go through a vastly higher number of iterations in a given time than their genetic counterparts. Thus, they were gradually shaped into better and better replicators by the – for the time being - still Darwinian forces of cultural evolution at a pace that left biological evolution panting behind (Dawkins, 1976). (Just as how in IT, innovations can occur much faster in software than in hardware.) Since all memes, irrespective of their symbiotic type, rely on cultural replication to spread, this creates competition among them, which in turn leads to an evolutionary arms race. Such a dynamic "researched and developed" new ways and scopes with which memes could be infectious (Dennett, 2017). The most useful ones, for our purposes, proved to be infectious by serving as "thinking tools" with which new chunks of the "design space" of possible memes could be opened up (p. 294-301). Such examples include phonemes, the alphabet, tallying, numbers, geometry, etc. Some have led to the creation of various physical tools, like the abacus, wax tablets, and the personal computer, all of which further extended the possibilities of thinking and with it the emergence of novel kinds of cultural information to be transmitted. Each expansion in the human arsenal of thinking tools with which further kinds of memes could be created allowed for more comprehension and top-down control. By today, humans live in a world where the most successful memes were created fully intentionally by people- scientists, philosophers, artists, politicians, religious leaders, advertisers, and bloggers. They are the prototypical example of creativity and intelligent design. Nevertheless, as Dennett (2017) points out, these are still joined in competition for brain space not just with each other, but with memes that are "semi-intelligently designed, hemi-semi-demi-intelligently designed, and evolutionary designed […], such as fads and fashions, pronunciation shifts and buzzwords" (p. 331). Moreover, they are themselves only made possible by the complex scaffoldings of such unintelligently designed memes. Consequently, while cultural evolution has considerably de-Darwinized due to the success of its own products, fossil traces of its Darwinian origins are still everywhere to be seen.

In sum, both Bayesian inference and Popperian epistemology are part of the solution to the problem of how most human knowledge is generated. In this light, the former framework can constitute a helpful extension to Deutsch's RTK. Human knowledge requires hierarchical Bayesian networks for hardware that are competent at picking up affordances from the environment. Animal brains were equipped with such networks as a result of genetic evolution. And, it requires memes for software, which are special kinds of affordances containing knowledge that was designed elsewhere, by cultural evolution, and which thus initiated a runaway process of layers and layers of cognitive competences being installed on top of one another in our brain. Dennett (2017) likens the event in which these two mechanisms joined together akin to how the eukaryotic cell - according to the endosymbiotic theory - came into existence as a result of an accidental merger between a bacterium and an archaeon. This was also an instance of a "technology transfer"; whereby the separately evolved competences of two entities were united in a fortuitous case of symbiosis that ushered in a huge leap in functionality (p. 389). We don't know how long it took for evolution by natural selection to iron out the details of the merger that resulted in a fully functioning eukaryotic cell. Presumably, countless iterations were wiped out whose configuration was not as stable as the one from which all subsequent multi-cellular life is descendant from today. In the same vein, we can only guess how the co-evolutionary process of Bayesian brains and memetic virtual machines got jumpstarted. Currently, we have but pieces of the puzzle. Trying to fill in the remaining gaps serves us with an exciting target of future research.

6. Concluding Remarks

In my thesis I made a case for what I termed the replicator theory of knowledge, which has its origins in the writings of David Deutsch. After examining the convergences between several explanatory theories from which this view emerges, I offered an explicit restatement of the theory, and outlined what we can gain by taking it seriously. I also argued that one of the main challenges to the theory's widespread legitimacy is the rise of Bayesianism in the cognitive

sciences. I then surveyed existing retorts to this contest, including that of Deutsch. Having found all of them to be lacking in some regards, I put forward my own preferred solution. In the context of this compromise position, I claimed that RTK has room for both Bayesian and Popperian epistemology.

By way of departing thoughts, I would like to flag two limitations that this current contribution has. Firstly, while in its scope this paper was constrained to only deal with the Bayesian challenge, there are other points of possible contest. For instance, the concept of replicators as being fundamental in evolutionary biology has been disputed (see Griffiths and Russell, 1994, Godfrey-Smith, 2000; Bourrat 2014). Its implications for cultural evolution even more so; there being an especially strong resistance to the terminology of memes (see Kuper, 2000; Atran, 2001; Boyd and Richerson, 2000; Richerson and Boyd, 2004). Besides limits of space, the decision to not extend the paper's analysis to these issues was motivated by the considerations that in the case of genetic evolution, despite these criticisms, Dawkins' Neo-Darwinism by and large remains the orthodoxy, and that in the case of cultural evolution, the pushback against memes, while substantial, has been already addressed by its defenders many times (see Dawkins, 1982, Dennett, 1995, 2017; Blackmore, 2000, 2006; Deutsch, 2011). Conversely, there has been no scholarly contribution that I am aware of that is dedicated to examining what the success of Bayesianism implies for the concept of replicators. Nevertheless, the future viability of RTK will depend on whether it can collectively fight off all these sources of problems.

Secondly, specifically the compromise position I offered as a possible way to reconcile Bayesianism and RTK retains for the time being a considerably speculative element due to lack of enough scientific evidence about the origins of language and culture. Empirical support is especially needed in finding out how exactly the human brain was able to pick up the earliest pieces of cultural information, and how those, in turn, shaped its underlying architecture, which launched the process of gene-culture coevolution. Of course, as we zoom in on the details of this problem to a larger extent, it may turn out that the position I advocated is untenable for

some reason or another. In fact, it may provide evidence against the notion of replicators all together. Regardless, as things currently stand, I believe that the unique take of RTK has valuable insights to offer for philosophers of knowledge.

Bibliography

Adriaans, P. (2020). Information. Stanford Encyclopedia of Philosophy. Available at: https://plato.stanford.edu/entries/information/

Armstrong, D. M. (1973). Belief, Truth, and Knowledge. Cambridge: Cambridge University Press.

Atran, S. (2001). The trouble with memes: Inference versus imitation in cultural creation. Human Nature, 12(4): 351-381.

Ayala, F. J. (2006): Evolution. Encyclopedia Britannica. Available at: https://www.britannica.com/science/evolution-scientific-theory

Bacon, F. (1620). Novum Organum. Thomas Fowler (ed., 1878). Oxford: Clarendon Press.

Beggs, E. J., and J.V. Tucker (2007). Can Newtonian Systems, Bounded in Space, Time, Mass and Energy Compute all Functions? Theoretical Computer Science, 371(1–2): 4–19.

Berkes, P., Orban, G., Lengyel, M., and Fiser, J. (2011). Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. Science, 331(6013): 83-7.

Blackmore, S. (2000). The Meme Machine. Oxford: Oxford University Press.

Blackmore, S. (2006). Memetics by Another Name. BioScience, 56(1): 74-75.

Bourrat, P. (2014). “From Survivors to Replicators: Evolution by Natural Selection Revisited”. Biology and Philosophy, 29(4): 517–538.

Boyd, R. and Richerson, P. (2000). Meme Theory Oversimplifies Cultural Change. Scientific American, 283(4): 55.

Brockman, J. (2019). Possible Minds: Twenty-Five Ways of Looking at AI. New York: Penguin Press.

Campbell, D. T. (1974). Evolutionary Epistemology. In The Philosophy of Karl Popper, Schlipp, P. A. (eds). La Salle, Illinois: The Open Court Publishing Company

Carey, S. (1985). Conceptual change in childhood. Cambridge: MIT Press.

Carter J. A. et al. (2017). Knowledge First. Oxford: Oxford University Press.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behavioral and brain sciences, 36(3): 181-204

Copeland, B. J. (2004). The Essential Turing: Seminal Writings in Computing, Logic, Philosophy, Artificial Intelligence, and Artificial Life. Oxford: Oxford University Press.

Copeland, B. J. (2018). The Church-Turing Thesis. Stanford Encyclopedia of Philosophy. Available at: https://plato.stanford.edu/entries/church-turing/

Creath, R. (2017). Logical Empiricism. Stanford Encyclopedia of Philosophy. Available at: https://plato.stanford.edu/entries/logical-empiricism/

Crupi, V. and Tentori, K. (2014). State of the field: Measuring information and confirmation. Studies in History and Philosophy of Science. 47: 81-90

Davidson, H. R. E. (1964). Gods and Myths of Northern Europe. New York: Penguin Books.

.

Dawkins, R. (1976). The Selfish Gene. Oxford: Oxford University Press.

Dawkins, R. (1982). The Extended Phenotype. Oxford: Oxford University Press.

Dawkins, R. (1983). Universal Darwinism. In: Evolution from molecules to man, Bendall, D. S. (ed.). Cambridge University Press.

De Mol, L. (2018). Turing Machines, Stanford Encyclopedia of Philosophy. Available at: https://plato.stanford.edu/entries/turing-machine/

Dennett, D. (1995). Darwin's Dangerous Idea. London: Penguin Press.

Dennett. D. (2017). From Bacteria to Bach and Back. London: W. W. Norton & Company.

DeRose, K. (2009). The Case for Contextualism, New York: Oxford University Press.

Deutsch, D. (1985). Quantum Theory, the Church-Turing Principle and the Universal Quantum Computer. Proceedings of the Royal Society of London A, 400: 97–117.

Deutsch, D. (1997). The Fabric of Reality. London: Penguin Press.

Deutsch, D. (2011). The Beginning of Infinity. London: Penguin Press.

Deutsch, D. (2012). Creative blocks, Aeon. Available at: https://aeon.co/essays/how-close-are-we-to-creating-artificial-intelligence

Deutsch, D. (2019). Beyond Reward and Punishment. In Brockman, J. (2019)

Donovan, T. M. and Mickey, R. M. (2019). Bayesian Statistics for Beginners: A Step-by-Step Approach. Oxford: Oxford University Press

Dorling, J. and Miller, D. (1981). Bayesian Personalism, Falsificationism, and the Problem of Induction, Proceedings of the Aristotelian Society, Supplementary Volumes, 55: 109-141

Dretske, Fred, 1981, Knowledge and the Flow of Information, Cambridge, MA: The MIT Press.

Dretske, F. (1985). “Precis of Knowledge and the Flow of Information”, in Naturalizing Epistemology, Kornblith, H. (ed.). Cambridge, MA: MIT Press: 169–187.

Duignan, B. (2009). Empiricism. Encyclopedia Britannica. Available at: https://www.britannica.com/topic/empiricism

Duhem, P. (1954). Aim and Structure of Physical Theories, trs. Philip Weiner. Princeton, NJ: Princeton University Press.

Dutant, J. (2015). The Legend of the Justified True Belief Analysis. Philosophical Perspectives 29 (1):95-145.

Elton, D. C. (2021). Applying Deutsch's concept of good explanations to artificial intelligence and neuroscience – An initial exploration. Cognitive Systems Research, 67: 9-17

Fisher, R. A. (1930). The Genetical Theory of Natural Selection. Clarendon Press, Oxford.

Floridi, L. (2010). Information: A Very Short Introduction. Oxford: Oxford University Press

Floridi, L. (2015). Semantic Conceptions of Information. Stanford Encyclopedia of Philosophy. Available at: https://plato.stanford.edu/entries/information-semantic/

Friston, K. J. (2017). Active Inference, Curiosity and Insight. Neural Computation, 29(10): 2633–2683.

Gettier, E. L. (1963). Is Justified True Belief Knowledge? Analysis, 23(6): 121–123.

Gleick, J. (2011). Information: A History, A Theory, A flood. New York: Pantheon Books.

Griffiths, P. E. and Russell D. G. (1994). Replicators and Vehicles? Or Developmental Systems". Behavioral and Brain Sciences, 17(4): 623–624.

Griffiths, T. L. and Tenenbaum, J. B. (2009). Theory-Based Causal Induction. Psychological Review, 116(4): 661–716

Godfrey-Smith, P. (2000). The Replicator in Retrospect. Biology and Philosophy, 15(3): 403–423.

Goertzel, B. (2020). The unorthodox path to AGI. Interview on TDS podcast by Harris, J. Available at: https://towardsdatascience.com/the-unorthodox-path-to-agi-a2fb633ca282

Goldman, A. I. (1967). A Causal Theory of Knowing. The Journal of Philosophy, 64(12): 357-372.

Goodfellow, I. and Bengio, Y. and Courville, A. (2016). Deep Learning. Cambridge, MA: MIT Press

Gopnik A. (1988). Conceptual and semantic development as theory change: The case of object permanence. Mind & Language, 3(3):197–216.

Gopnik, A. and Tenenbaum, J. B. (2007). Bayesian networks, Bayesian learning and cognitive development. Developmental Science, 10(3):281–287.

Gopnik, A. and Wellman, H. M. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms and the theory theory. Psychological Bulletin, 138(6): 1085–1108.

Hahn, U. (2014). The Bayesian boom: good thing or bad? Frontiers in Psychology, 5:765.

Haldane, J. B. S. (1924-1934). A Mathematical Theory of Natural and Artificial Selection. A series of papers in Transactions of the Cambridge Philosophical Society 23-28 and in Genetics 19

Hamilton W. D. (1964). The Genetical Evolution of Social Behaviour. I and II". Journal of Theoretical Biology, 7(1): 1–16 and 17-52

Henderson, L. (2018). The Problem of Induction. Stanford Encyclopedia of Philosophy. Available at: https://plato.stanford.edu/entries/induction-problem/

Howson, C. and Urbach, P. (2006). Scientific Reasoning: The Bayesian Approach. Chicago: Open Court Publishing.

Hume, D. (1748). An Enquiry Concerning Human Understanding, Oxford: Oxford University Press.

Hutter, M. (2000). A Theory of Universal Artificial Intelligence based on Algorithmic Complexity. arXiv:cs/0004001

Ichikawa, J. J. (20179. The Analysis of Knowledge. Stanford Encyclopedia of Philosophy. Available at: https://plato.stanford.edu/entries/knowledge-analysis/

Jeffrey, R. C. (1975). Probability and Falsification: Critique of the Popper Program. Synthese, 30(1/2): 95–117.

Kahneman, D. (2011). Thinking, Fast and Slow. New York: Farrar, Straus and Giroux

Kant, I. (1781/1998). Critique of Pure Reason. Cambridge University Press.

Kuper A. (2000). If Memes Are the Answer, What Is the Question? in Darwinizing Culture, R. Aunger (ed.). Oxford: Oxford University Press. pp. 175–188.

Largent, M. A. (2009). The So-Called Eclipse of Darwinism. Transactions of the American Philosophical Society, 99(1): 3-21

Levy, S. (1977). Defeasibility Theories of Knowledge. Canadian Journal of Philosophy, 7(1): 115-123.

Locke, J. (1689). An Essay Concerning Human Understanding. Peter H. Nidditch (ed.), 1975. Oxford: Clarendon Press

Longion, H. (2016). Underdetermination: A Dirty Little Secret? Issue 4 of STS occasional papers. Available at: https://www.ucl.ac.uk/sts/sites/sts/files/longino_2016_underdetermination.pdf

Magee, B. (1973). Popper. London: Penguin.

McGrayne, S. B. (2011). The Theory that Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy. New Haven: Yale University Press

Miller, D. (1994). Critical Rationalism: A Restatement and Defence, Chicago: Open Court.

Milne, P. (1995). A Bayesian Defence of Popperian Science? Analysis, 55(3): 213-215.

Németi, I., and G. Dávid (2006). Relativistic Computers and the Turing Barrier. Journal of Applied Mathematics and Computation, 178(1): 118–142.

Newton, I. (1713). Philosophiae Naturalis Principia Mathematica, General Scholium. Second edition, I. Bernard Cohen and Anne Whitman's 1999 translation, University of California Press

Okasha, S. (2002). Philosophy of Science: A Very Short Introduction. Oxford: Oxford University Press.

Paley, F. A. (2012). The Theaetetus of Plato: Translated, With Introduction and Brief Explanatory Notes. London: Forgotten Books

Papineau, D. (2019). Knowledge is crude. Aeon. Available at: https://aeon.co/essays/knowledge-is-a-stone-age-concept-were-better-off-without-it

Piccinini, G. (2017). Computation in Physical Systems. Stanford Encyclopedia of Philosophy. Available at: https://plato.stanford.edu/entries/computation-physicalsystems/

Pitowsky, I. (1990). The Physical Church Thesis and Physical Computational Complexity. Iyyun, 39: 81–99.

Popper, K. (1959). The Logic of Scientific Discovery, translation by the author of Logik der Forschung (1935), London: Hutchinson.

Popper, K. (1963). Conjectures and Refutations. London: Routledge.

Popper, K. (1976). Unended Quest: An Intellectual Autobiography: London: Fontana.

Popper, K (1979). Objective Knowledge: An Evolutionary Approach. Oxford: Clarendon Press.

Popper, K. (1994). The Myth of the Framework: In Defense of Science and Rationality. London: Routledge

Ray, M. (2017). Lamarckism. Encyclopedia Britannica. Available at: https://www.britannica.com/science/Lamarckism

Richerson, P. and Boyd, R. (2004), Not by Genes Alone: How Culture Transformed Human Evolution, Chicago: University of Chicago Press.

Rosenkrantz, R. D. (1977). Inference, Method and Decision: Towards a Bayesian Philosophy of Science. Boston: D. Reidel Publishing Company

Russell, B. (1946). A History of Western Philosophy. London: George Allen and Unwin Ltd.

Russell, B. (1948). Human Knowledge: Its Scope and Limits. New York: Simon and Schuster.

Sellars, W. (1962). Philosophy and the Scientific Image of Man, in Frontiers of Science and Philosophy, Robert Colodny (ed.). Pittsburgh: University of Pittsburgh Press

Shannon, C. E. (1948). A Mathematical Theory of Communication. Bell System Technical Journal, 27(3): 379–423 & 27(4): 623–656.

Shannon, C. E. and Weaver, N. (1949). The Mathematical Theory of Communication. Urbana and Chicago: University of Illinois Press

Stanley, J. (2005). Knowledge and Practical Interests. New York: Oxford University Press.

Tegmark, M. (2017). Life 3.0: Being Human in the Age of Artificial Intelligence. New York: Penguin Random House

Tenenbaum et al. (2011). How to Grow a Mind: Statistics, Structure, and Abstraction. Science, 331(6022): 1279-1285.

Tomasello, M. (2014). A Natural History of Human Thinking. Cambridge, MA: Harvard University Press

Turing, A. M. (1936). On Computable Numbers, with an Application to the Entscheidungsproblem. Proceedings of the London Mathematical Society, 2(42): 230–265

Turing, A. M. (1948). Intelligent Machinery, National Physical Laboratory Report, in Copeland 2004b: 410–432.

Turing, A. M. (1950). Computing Machinery and Intelligence. Mind, 59(236): 433–460

von Neumann, J. (1948). The General and Logical Theory of Automata. In John von Neumann: Collected Works. Volume 5. Taub, A. H. (eds., 1963).

Wellman, H. M. (1990) The child's theory of mind. Cambridge: MIT Press.

Williams, G. C. (1966). Adaptation and Natural Selection. Princeton: Princeton University Press.

Williamson, T. (2000). Knowledge and its Limits, Oxford: Oxford University Press.

Yudkowsky, E. S. (2006). An Intuitive Explanation of Bayesian Reasoning. Available at: https://www.yudkowsky.net/rational/bayes

Zagzebski, L. (1994). “The Inescapability of Gettier Problems”, The Philosophical Quarterly, 44(174): 65–73.