PUTNAM'S DIAGONAL ARGUMENT AND THE IMPOSSIBILITY OF A UNIVERSAL LEARNING MACHINE

TOM F. STERKENBURG

ABSTRACT. Putnam (1963) construed the aim of Carnap's program of inductive logic as the specification of a "universal learning machine," and presented a diagonal proof against the very possibility of such a thing. Yet the ideas of Solomonoff (1964) and Levin (1970) lead to a mathematical foundation of precisely those aspects of Carnap's program that Putnam took issue with, and in particular, resurrect the notion of a universal mechanical rule for induction.

In this paper, I take up the question whether the Solomonoff-Levin proposal is successful in this respect. I expose the general strategy to evade Putnam's argument, leading to a broader discussion of the outer limits of mechanized induction. I argue that this strategy ultimately still succumbs to diagonalization, reinforcing Putnam's impossibility claim.

1. INTRODUCTION

Putnam (1963a) famously challenged the feasibility of Carnap's program of inductive logic on the grounds that a quantitative definition of "degree of confirmation" can never be adequate as a rational reconstruction of inductive reasoning. Specifically, he formulated two conditions of adequacy on any reconstruction of "the judgements an ideal inductive judge would make" (ibid., 778), and proceeded to give a *diagonal proof* to the effect that no Carnapian measure function can satisfy both. In (1963b), Putnam explicitly assumed the view that "the task of inductive logic is the construction of a 'universal learning machine'" (ibid., 303), and accordingly presented his proof as showing the impossibility of this notion. What was shown, in these terms, is that there can be no *learning machine* that is also *universal*: no inductive method that is effectively computable, that is also able to eventually detect any pattern that is effectively computable.

Independently of the work of Putnam, the suggestions of Solomonoff (1964) towards an "optimum induction system" gave rise to a definition that is very much in this spirit. The elements that Solomonoff took from Carnap's program, and those that he added to it—most importantly, the central role of effective computability—are the very elements that Putnam presumed in his challenge to it. Solomonoff's ideas found a secure mathematical footing in the work by Levin (1970), resulting in what qualifies, perhaps, as the definition of a universal inductive machine (also see Li and Vitányi, 2008; Hutter, 2007; Rathmanner and Hutter, 2011). Namely, the

Date: February 7, 2018.

This is the final version, as accepted for publication in *Erkenntnis* (doi: 10.1007/s10670-018-9975-x). This paper was written while I was with the Machine Learning group, Centrum Wiskunde & Informatica, Amsterdam, and the Faculty of Philosophy, University of Groningen. I want to thank Peter Grünwald, Wouter Koolen, Jan Leike, and Nishant Mehta for helpful discussions, Jan-Willem Romeijn for valuable advice on earlier versions of the paper, and finally the anonymous reviewers for their careful comments, which did much to improve it.

Solomonoff-Levin measure does manage to unite versions of Putnam's two adequacy conditions—though, crucially, involving a weakened notion of effective computability.

In this paper I investigate whether the Solomonoff-Levin proposal indeed gives a definition of an "optimum," "cleverest possible," or *universal* inductive machine. More broadly, this is an investigation into the possibility of a perfectly general and purely mechanical rule for extrapolating data—against the lesson that has generally been taken from Putnam that "[t]here is no universal algorithm" for induction (Dawid, 1985b, 341; also see van Fraassen, 2000, 260; 1989, 132ff). I will argue that there is promise in the general strategy that underlies the Solomonoff-Levin proposal, which is to try and identify a natural class of effective elements that is immune to diagonalization. This opens the prospect of attaining plausible versions of Putnam's two conditions that *are* compatible, and that enable a notion of an inductive rule that is universal in a Reichenbachian sense: this *optimal* inductive rule will learn successfully if any inductive rule does. I will then show, however, that Putnam's lesson prevails: on a closer inspection of the proper interpretation of the relevant elements we see that this general strategy cannot escape diagonalization after all.

2. Overview

First, in section 3, I will introduce Putnam's original argument, which shows that no confirmation function can fulfill both of two conditions to qualify as a universal inductive rule: the first on its convergence to any effectively computable hypothesis, the second on it being effectively computable itself. This is only one part of Putnam's charge; the other is that this is a defect peculiar to confirmation functions, because other methods, that respect the role of scientific theories (in particular, the hypothetico-deductive or HD method), *can* satisfy it. Next, in section 4, I explain how Solomonoff took his cue from Carnap's project, and went on to develop his ideas in a direction that (perhaps unlike Carnap's own approach) falls squarely within the general outlook and formal set-up that Putnam assumed for his argument. This raises the question how the resulting Solomonoff-Levin definition evades the diagonal argument.

The only way around Putnam's argument is to argue for a weakening of at least one of the two conditions that he showed are incompatible. Hence the question is what weakening the Solomonoff-Levin proposal introduces, and whether it can be given a proper motivation. To be in a position to answer this question, I need to go through a more technical exposition, section 5, that traces the way to the exact definition of the Solomonoff-Levin measure function. Here I describe the general strategy of identifying a class of effective measure functions that cannot be diagonalized; the Solomonoff-Levin measure is a universal element in this class. This definition does satisfy Putnam's first condition on convergence to any computable hypothesis, but it is effective in too weak a sense to still satisfy Putnam's second condition.

Turning to the question whether an accordingly weakened condition is defensible, I must first consider the second component of Putnam's charge. This is the claim that the conjunction of the two original conditions is not unreasonably strong, since the HD procedure does satisfy it. The conclusion that I reach in section 6 is that this claim does not stand up to scrutiny: drawing a distinction between specific *methods* and an underlying *architecture*, we see that the HD approach and the Bayesian approach of confirmation functions are in the same predicament. The lesson of the diagonal argument is rather that no fully specified or *fixed* method can satisfy this pair of conditions. Given this, it stands to reason to explore the possibility of a notion of a universal inductive rule that only satisfies a weaker pair. This I do in the final part of the paper, through an evaluation of the Solomonoff-Levin proposal.

I start in section 7 with the question whether the Solomonoff-Levin function, in the spirit of the first condition, can detect all reasonable patterns. The naive interpretation of this question fails to be convincing, which prompts a different and much more natural interpretation. This Reichenbachian interpretation, pursued in section 8, takes the Solomonoff-Levin functions as *optimal* among all possible inductive rules. If the original class of effective measure functions represents all possible inductive rules, then the Solomonoff-Levin measure, as a universal element, is in a precise sense at least as good as any possible inductive rule. In general, the identification of an undiagonalizable class of elements, if conjoined with a successful argument that it represents all possible inductive rules, yields a notion of a universal inductive rule.

Unfortunately, there is a problem with this strategy, a problem that even goes beyond worries about the weaker notion of computable approximability. Namely, it is obstructed by the fact that inductive rules should actually be identified with confirmation functions, i.e., *conditional* measure functions. This fact might sound innocuous, but it impacts their effectiveness properties. I will show that Putnam's original argument implies that this indeed blocks the central strategy of identifying a class of effective elements that cannot be diagonalized. Thus, as I conclude in section 9, the analysis of this paper provides further support to Putnam's case: there can be no such thing as a universal inductive rule.

3. Putnam's argument

Consider a simple first-order language with a single monadic predicate G and an ordered infinity of individuals x_i, $i \in \mathbb{N}^{>0}$. Let a *computable hypothesis* h be a computable set of sentences $h(x_i)$ for each individual x_i, where $h(x_i)$ equals one of Gx_i and $\neg Gx_i$. A Carnapian *confirmation function* C gives the degree of confirmation that one statement confers upon another. In particular,

$$\mathrm{C}(h(x_{n+1}), h(x_1) \ \& \ \dots \ \& \ h(x_n))$$

is the degree to which the statement that the next individual x_{n+1} satisfies h is confirmed by the fact that all of x_1 up to x_n do so already. (Carnap also calls this the *instance confirmation* of h.) Now, if a given Carnapian confirmation function is supposed to be a rational reconstruction of our inductive practice, then, since our actual inductive methods would be sure to discern any computable pattern eventually, so should this given confirmation function. Hence a condition of adequacy on such a confirmation function C is that

(I) For any computable hypothesis h, the value for the instance confirmation $\mathrm{C}(h(x_{n+1}), h(x_1) \ \& \ \dots \ \& \ h(x_n))$ should converge to 1 as we observe a longer and longer succession of confirming individuals $x_1, \dots, x_n$.

But for any confirmation function C that itself satisfies a weak condition of effective computability (to not be "of no use to anybody," Putnam, 1963a, 768):

(II) For every n, it must be possible to compute a k such that if G holds for the next k individuals $x_{n+1}, \ldots, x_{n+k}$, then the instance confirmation $\mathrm{C}\left(Gx_{n+k+1}, Gx_{n+1} \ \& \ \ldots \ \& \ G(x_{n+k})\right)$ exceeds 0.5,

one can prove by diagonalization C's violation of (I). This is Putnam's diagonal argument: if the ideal inductive policy is to fulfill (I) and (II), then it is provably impossible to reconstruct it as a Carnapian confirmation function.

Let me simplify things a little. We can treat condition (I) as an instance of the condition on an 'inductive method' M, a condition I will leave slightly informal in its generality, that

(I*) M converges on any true computable hypothesis.

Moreover, in later expositions of the argument (e.g., Earman, 1992, 207ff; Kelly, 2004, 701f), the somewhat cumbersome condition (II) is often replaced by the (stronger) condition that C is simply a computable function. The general condition on an inductive method M is that

(II*) M is computable.

The diagonal proof of the incompatiblity of (I*) and (II*) for confirmation functions is straightforward. Given candidate computable confirmation function C, we construct a computable hypothesis h such that C fails to converge on h, as follows. Starting with the first individual x_1, compute $\mathrm{C}(Gx_1, \top)$ and let $h(x_1)$ be $\neg Gx_1$ precisely if $\mathrm{C}(Gx_1, \top) > 0.5$. For each new individual x_{n+1}, proceed in the same fashion: compute $\mathrm{C}(Gx_{n+1}, h(x_1) \ \& \ \ldots \ \& \ h(x_n))$ and let $h(x_{n+1})$ be $\neg Gx_{n+1}$ precisely if this probability is greater than 0.5. The hypothesis h is clearly computable, but by construction the instance confirmation given by C does not converge to 1: indeed, it never even goes above 0.5. Thus, again, if the ideal inductive policy is to be able to converge to any computable hypothesis, *and* is to be computable itself, then it is impossible to reconstruct it as a confirmation function.

But maybe such a policy is so idealized as to escape any formalization? To seal the fate of Carnap's program, Putnam proceeds to give an example of an inductive method that is *not* based on a confirmation function and that *does* satisfy the two requirements. This method HD is the *hypothetico-deductive method*: supposing some enumeration of hypotheses that are proposed over time, at each point in time select and use for prediction (*accept*) the hypothesis first in line among those that have been consistent with past data. Then it satisfies convergence condition (I*), or more precisely:

($\mathrm{I}^\dagger$) For any true computable hypothesis h, if h is ever proposed, then HD will eventually come to (and forever remain to) accept it.

The distinctive feature of HD is that it relies on the hypotheses that are actually proposed. To Putnam, this is as it should be. Not only does it conform to scientific practice: more fundamentally, it does justice to the "*indispensability of theories* as instruments of prediction" (ibid., 778). This appears to be the overarching reason why Putnam takes issue with Carnap's program (ibid., 780):

> Certainly it appears implausible to say that there is a *rule* whereby one can go from the observational facts ... to the observational prediction without any 'detour' into the realm of theory. But this is a consequence of the supposition that degree of confirmation can be "adequately defined"; i.e. defined in such a way as to agree with the actual inductive judgements of good and careful scientists.

Incredulously (ibid., 781):

> . . . we get the further consequence that it is possible in principle to build an electronic computer such that, if it could somehow be given all the observational facts, it would always make the best prediction—i.e. the prediction that would be made by the best possible scientist if he had the best possible theories. *Science could in principle be done by a moron* (or an electronic computer).

Here Putnam is still careful not to attribute to Carnap too strong a view: "Of course, I am not accusing Carnap of believing or stating that such a rule exists; the existence of such a rule is a *disguised* consequence of the assumption that [degree of confirmation] can be 'adequately defined'" (ibid., 780). Nevertheless, in his *Radio Free Europe* address (1963b), Putnam declares that "we may think of a system of inductive logic as a design for a 'learning machine': that is to say, a design for a computing machine that can extrapolate certain kinds of empirical regularities from the data with which it is supplied" (ibid., 297); and "if there is such a thing as a correct 'degree of confirmation' which can be fixed once and for all, then a machine which predicted in accordance with the degree of confirmation would be an *optimal*, that is to say, a cleverest possible learning machine" (ibid., 298). Again, the diagonal proof would show that there can be no such thing: it is "an argument against the existence – that is, against the possible existence – of a 'cleverest possible' learning machine" (ibid., 299).

4. Solomonoff's new start

Solomonoff (1964) aimed to describe precisely that: an "optimum" inductive method, a formal system of inductive inference that "is at least as good as any other that may be proposed" (ibid., 5). His ideas can indeed be seen as a particular offspring of Carnap's inductive logic; one that takes Putnam's picture of a learning machine seriously.

Solomonoff's objective is clear (ibid., 2):

> The problem dealt with will be the extrapolation of a long sequence of symbols—these symbols being drawn from some finite alphabet. More specifically, given a long sequence, represented by T, what is the probability that it will be followed by the subsequence represented by a? In the language of Carnap (1950), we want $c(a, T)$, the degree of confirmation of the hypothesis that a will follow, given the evidence that T has just occurred.

The underlying motivation is also very much in accord with things Carnap writes in his 1950 book. Solomonoff's suggestion that "all problems in inductive inference . . . can be expressed in the form of the extrapolation of a long sequence of symbols" (ibid.) parallels Carnap's insistence on the primacy of the predictive inference—"the most important and fundamental inductive inference" (1950, 207). Carnap's *requirement of total evidence* (see ibid., 211ff) returns in Solomonoff's remark that "the corpus that we will extrapolate . . . *must contain all of the information that we want to use in the induction*" (ibid., 8). And Carnap's discussion under the header "Are Laws Needed for Making Predictions?" (ibid., 574f)—conclusion: "the use of laws is not indispensable"—is easily read as informing Solomonoff's statement that his proposed methods are "meant to bypass the explicit formulation of scientific laws, and use the data of the past directly to make inductive inferences about specific future events" (1964, 16).

This already very much resembles the picture that Putnam painted in order to challenge it. What is more, the problem setting of sequence extrapolation is readily translatable into the formal set-up that Putnam presupposes in his paper. Let us suppose, as is customary in modern discussions of Solomonoff's theory, that we have an alphabet of only two symbols, '0' and '1.' Now Putnam assumes with Carnap a monadic predicate language L, but with an *ordered* domain $x_1, x_2, x_3, \dots$ of individuals. Let L have a single monadic predicate G. Identifying the individuals with positions in a sequence as Putnam does (1963a, 766), we can have a '1' at the i-th position express the fact that individual x_i satisfies G, and a '0' that it does not. Thus we translate a symbol sequence of length t into the observation of the first t individuals.

Solomonoff's setting is then fully within the scope of Putnam's argument. This in contrast to that of Carnap, who could still resort to the defense that in his works he does *not* assume an ordered domain, and so "the difficulties which Putnam discusses do not apply to the inductive methods which I have presented in my publications" (1963a, 986). Nevertheless, Carnap does acknowledge at various places the need for taking into consideration the order of individuals in explicating degree of confirmation (e.g., 1950, 62ff; 1963b, 225f); and he envisioned for this future project the same kind of "coordinate language" that Putnam assumes (also see Skyrms, 1991). For such a language, Carnap should have agreed with Putnam's charge that an inductive system that is "not 'clever' enough to learn that position in the sequence is relevant" is too weak to be adequate. The difference in opinion then ultimately comes down to *what* regularities in the observed individuals should be extrapolated (i.e., *what* hypotheses or patterns should gain higher instance confirmation from supporting observations).

Carnap states in (1963a, 987; 1963b, 226) that he would only consider "laws of finite span." In terms of symbol sequence extrapolation, these are the hypotheses that make the probability of a certain symbol's occurrence at a certain position only depend on the immediately preceding subsequence of a fixed finite length (i.e., a Markov chain of certain order). In particular, hypotheses must not refer to *absolute* coordinates, which immediately rules out Putnam's example of the hypothesis that "the prime numbers are occupied by red" (1963a, 765). In Carnap's view, "no physicist would seriously consider a law like Putnam's prime number law" (1963a, 987), hence "it is hardly worthwhile to take account of such laws in adequacy conditions for [confirmation functions]" (1963b, 226). According to Putnam, however, "existing inductive methods are capable of establishing the correctness of such a hypothesis ... and so must any adequate 'reconstruction' of these methods" (1963a, 765). Indeed, the same goes for *any* effectively computable pattern; this is his adequacy condition (I*).

Others have charged Carnap's confirmation functions with an inability to meet various adequacy conditions on recognizing regularities (notably Achinstein, 1963; in fact the critique of Goodman, 1946, 1947 can be seen as an early instance of this line of attack). What is distinctive about Putnam's adequacy conditions is the emphasis on effective computability. Interestingly, this notion of effective computability is also the fundamental ingredient in Solomonoff's proposal. It is this aspect that genuinely sets Solomonoff's approach apart from Carnap's. The measure functions that Solomonoff proposed in (1964), and that evolved in the modern definition of a measure function Q_U that we will see below, were explicitly defined in terms of the inputs to a universal Turing machine. Moreover, one can show that the instance confirmation via Q_U of *any true computable hypothesis* will converge to 1, thus fulfilling convergence condition (I*).

5. The Solomonoff-Levin measure

How could Solomonoff evade Putnam's diagonal argument? If the Solomonoff-Levin function Q_U is within the scope of Putnam's argument, and it still fulfills convergence condition (I*), then it must give way with respect to effectiveness condition (II*). To explain how Q_U fulfills (I*) but not (II*), we will need to go into the details. This we do in the current section; in the next section we return to the main thread and see what this means for Q_U as a purported "optimum," or *universal* inductive rule.

Specifically, we will work in this section towards the precise specification of Q_U, and show that it satisfies (I*). For a large part this amounts to retracing the formal setting that was developed in the landmark paper of Zvonkin and Levin (1970), based on Levin's doctoral thesis (translated as Levin, 2010).

We start with the notion of a computable (probability) measure on the Cantor space $\{0,1\}^\omega$, the set of all infinite sequences of symbols in $\{0,1\}$. More accurately, a measure on Cantor space is defined on a tuple $(\{0,1\}^\omega, \mathfrak{F})$, with $\mathfrak{F}$ a σ-algebra on $\{0,1\}^\omega$. Then a probability measure on $(\{0,1\}^\omega, \mathfrak{F})$ is a countably additive function $\mu : \mathfrak{F} \to [0,1]$ with $\mu(\{0,1\}^\omega) = 1$. Let the *basic cylinder* $[\![\boldsymbol{x}]\!]$ be the class of all infinite extensions in $\{0,1\}^\omega$ of the *finite* sequence $\boldsymbol{x} \in \{0,1\}^*$. It is convenient to view a measure (as well as the associated σ-algebra $\mathfrak{F}$) as being generated from an assignment of probability values to just the basic cylinders $[\![\boldsymbol{x}]\!]$ for all finite sequences $\boldsymbol{x}$. That is, we view a measure as being generated from a *pre-measure*, a function $m : \mathbb{B}^* \to [0,1]$ on the finite sequences that satisfies $m(\boldsymbol{\varnothing}) = 1$ for the *empty* sequence $\boldsymbol{\varnothing}$ and $m(\boldsymbol{x}0)+m(\boldsymbol{x}1) = m(\boldsymbol{x})$ for all $\boldsymbol{x} \in \{0,1\}^*$ and its one-symbol extensions $\boldsymbol{x}0$ and $\boldsymbol{x}1$. The extension theorem due to Carathéodory (see Tao, 2011, 148ff) then gives a σ-algebra $\mathfrak{F}$ over $\{0,1\}^\omega$ (which includes all Borel classes) and unique measure μ_m on $\mathfrak{F}$ with $\mu_m([\![\boldsymbol{x}]\!]) = m(\boldsymbol{x})$. I will follow the custom of simply writing '$\mu(\boldsymbol{x})$' for '$\mu([\![\boldsymbol{x}]\!])$.' (See Reimann, 2009, 249ff; Nies, 2009, 68ff for more details.)

The most basic example of a measure on Cantor space is the *uniform* measure λ. It is generated from the pre-measure with $m(\boldsymbol{x}) = 2^{-|\boldsymbol{x}|}$ for all $\boldsymbol{x}$, where $|\boldsymbol{x}|$ denotes $\boldsymbol{x}$'s length.

Now a measure is *computable* if it is generated from a computable pre-measure. A pre-measure m is computable if its values can be uniformly computed up to any given precision. That is, there is a computable $f : \{0,1\}^* \times \mathbb{N} \to \mathbb{Q}$ such that $|f(\boldsymbol{x}, s) - m(\boldsymbol{x})| < 2^{-s}$ for all $\boldsymbol{x} \in \{0,1\}^*, s \in \mathbb{N}$ (see Downey and Hirschfeldt, 2010, 202f). I will adopt the nomenclature of the *arithmetical hierarchy* of levels of effective computability (see Soare, 2016, 79ff) and henceforth refer to the computable measures as the Δ_1 ('delta-one') measures.

We will see below that the Solomonoff-Levin measure Q_U has the property that for any Δ_1 measure μ, if the data is in fact generated by μ, then with probability 1 ('μ-almost surely') the values $Q_U(x_{n+1} \mid \boldsymbol{x}^n) = \frac{Q_U(\boldsymbol{x}^{n+1})}{Q_U(\boldsymbol{x}^n)}$ for $x_{n+1} \in \{0,1\}, \boldsymbol{x}^n \in \{0,1\}^n$ converge to the values $\mu(x_{n+1} \mid \boldsymbol{x}^n)$ as n goes to infinity. That is, Q_U satisfies the following condition on an inductive method M:

(I: Δ_1) M converges μ-almost surely to Δ_1 measure μ.

This is an instance of convergence condition (I*) on a measure function, that at the same time generalizes from 'deterministic' computable hypotheses or single infinite computable sequences to hypotheses that are probability measures on infinite sequences. (The special case of a computable infinite sequence $\boldsymbol{x}^\omega$ corresponds to a Δ_1 measure that assigns probability 1 to every initial segment $\boldsymbol{x}^n$ of $\boldsymbol{x}^\omega$.) Moreover, we can rephrase effectiveness condition (II*) on a measure as

(II: Δ_1) M is Δ_1.

This condition is *not* satisfied by Q_U. It is effectively computable in a weaker sense, that we turn to now.

Namely, we proceed with the notion of a *semi-computable* or Σ_1 ('sigma-one') measure on the extended space $\{0,1\}^\omega \cup \{0,1\}^*$ of infinite *and finite* sequences. This notion will strike those who see it for the first time as cumbersome, if not downright awkward; I will try to explain in what sense it is both natural and important. First I will briefly describe how this class of measures comes about as precisely the *effective transformations* of the uniform measure on the Cantor space. Then I will discuss the crucial property of this class that *it cannot be diagonalized*, meaning that it contains *universal elements*. The Solomonoff-Levin measure is such a universal element.

Let a *transformation* λ_F of the uniform measure by Borel function $F : \{0,1\}^\omega \to \{0,1\}^\omega$ be defined by $\lambda_F(A) = \lambda(F^{-1}(A))$. Every Borel measure μ on Cantor space can be obtained as a transformation of λ by some Borel function (see Reimann, 2009, 252f).

We will now consider transformations by functions that are *effectively computable*. There are some details involved in the need to downscale these transformations to functions f on *finite* sequences, in order to impose the property of computability (see ibid., 253f); in the end we are led to precisely those functions that can be represented by a particular type of Turing machine. Originally dubbed an *algorithmic process* (Zvonkin and Levin, 1970, 99), this type of machine is now better known as a *monotone* machine (see Shen et al. 2017): it can be visualized as operating on a steady stream of input symbols, producing an (in)finite output sequence in the process. We then indeed have an effective analogue to the earlier statement: every Δ_1 measure can be obtained as a transformation λ_M of the uniform measure by some monotone machine M (Zvonkin and Levin, 1970, 100f).

The monotone machines leading to the Δ_1 measures have the special property that they are 'almost total,' meaning that they produce an unending sequence on λ-almost all infinite input streams (ibid.). In general, however, a monotone machine M can fail to do so. This translates into the possibility that $\lambda_M(\boldsymbol{x})$ is strictly greater than $\lambda_M(\boldsymbol{x}0) + \lambda_M(\boldsymbol{x}1)$ for some $\boldsymbol{x}$. In that case we can say that λ_M assigns positive probability to the *finite* sequence $\boldsymbol{x}$. A function λ_M can thus be interpreted as a measure on the collection of infinite *and* finite sequences.

Levin calls the class of transformations λ_M by all monotone machines M the class of *semi-computable* measures on $\{0,1\}^\omega \cup \{0,1\}^*$. This is because the pre-measures corresponding to these transformations are precisely the functions $m : \{0,1\}^* \to [0,1]$ with $m(\boldsymbol{x}) \geq m(\boldsymbol{x}0) + m(\boldsymbol{x}1)$ for all $\boldsymbol{x}$ that satisfy a weaker requirement of computability, that can be paraphrased as *computable approximability from below* (Zvonkin and Levin, 1970, 102f). In exact terms (see Downey and Hirschfeldt, 2010, 202f), we call m (lower) semi-computable if there is a computable $f : \{0,1\}^* \times \mathbb{N} \to \mathbb{Q}$ such that for all $\boldsymbol{x} \in \{0,1\}^*$ we have $f(\boldsymbol{x}, s) \leq f(\boldsymbol{x}, s+1)$ for all $s \in \mathbb{N}$ and

$\lim_{s\to\infty} f(\boldsymbol{x}, s) = m(\boldsymbol{x})$. Equivalently, the so-called *left-cut* $\{(q, \boldsymbol{x}) \in \mathbb{Q} \times \{0,1\}^* : q < m(\boldsymbol{x})\}$ is computably enumerable or Σ_1. For that reason I will refer to a semi-computable measure on $\{0,1\}^\omega \cup \{0,1\}^*$ as a Σ_1 measure.

Let me reiterate the parallel between, on the one hand, the expansion from the Δ_1 to the Σ_1 measures, and, on the other, the expansion from the *total* computable (t.c.) to the *partial* computable (p.c.) functions. It is well-known since Turing (1936) that the class of t.c. functions is diagonalizable, and that this is overcome by enlarging the class to the p.c. functions (see Soare, 2016, 4ff). More precisely: under the assumption that there exists a *universal* t.c. function $\mathring{f}$ that can emulate every other t.c. function (meaning that $\mathring{f}(i, x) = f_i(x)$ for a listing $\{f_i\}_{i\in\mathbb{N}}$ of all t.c. functions), we can directly infer a *diagonal function* g (say $g(x) := \mathring{f}(x, x) + 1$) that is t.c. yet distinct from every single f_i (because $g(i) = f_i(i) + 1 \neq f_i(i)$ for all i), which is a contradiction. (Note the similarity to the argument in section 3.) To say that the class of t.c. functions is diagonalizable is therefore to say that there can be no such universal $\mathring{f}$, hence no effective listing of all elements: *the class is not effectively enumerable.* The introduction of partialness, however, defeats the construction of a diagonal function (consider: what if $f_i(i)$ is undefined?); and indeed the class of p.c. functions *is* effectively enumerable, *does* contain universal elements. Likewise, the class of Δ_1 measures is not effectively enumerable, does not contain universal elements; the larger class of Σ_1 measures is and does. I now turn to these universal Σ_1 measures (Zvonkin and Levin, 1970, 103f).

Informally, a universal Σ_1 measure "is 'larger' than any other measure, and is concentrated on the widest subset of $\{0,1\}^\omega \cup \{0,1\}^*$" (ibid., 104, notation mine). Formally, a universal Σ_1 measure $\mathring{\mu}$ is such that it *majorizes* every other Σ_1 measure: for every $\mu_i \in \Sigma_1$ there is a constant $c_i \in [0, 1]$ such that for all $\boldsymbol{x} \in \{0,1\}^*$ it holds that $\mathring{\mu}(\boldsymbol{x}) \geq c_i \cdot \mu_i(\boldsymbol{x})$. "This fact is one of the reasons for introducing the concept of semi-computable measure" (ibid.)—we may take it as the main reason.

A natural way of obtaining a universal Σ_1 measure is the following. Since the monotone machines can also be effectively enumerated, we can likewise specify *universal* such machines. Let $\{\boldsymbol{z}_i\}_{i\in\mathbb{N}}$ be some prefix-free set of finite strings that serves as an encoding of all monotone machines M_e: the universal monotone machine that employs this encoding emulates any other monotone machine M_e on first receiving the corresponding code sequence $\boldsymbol{z}_e$. More precisely, this universal machine is such that it produces output $\boldsymbol{x}$ on input $\boldsymbol{z}_e\boldsymbol{y}$ if and only if M_e produces output $\boldsymbol{x}$ on input $\boldsymbol{y}$. Now a transformation λ_U of λ by such a universal machine U yields a universal Σ_1 measure.

We have finally arrived at the definition of the Solomonoff-Levin measure. The measure Q_U is precisely the transformation of λ by universal monotone machine U.

Definition 1. $Q_U := \lambda_U$.

So there are in fact infinitely many such measures, one for each choice of universal monotone machine U. Each is a universal Σ_1 measure. It is this property that is exploited in the adequacy result.

Proposition 2. Q_U fulfills (I: Δ_1).

Proof. Let μ be a Δ_1 measure. The fact that Q_U majorizes μ entails that μ is absolutely continuous with respect to Q_U (i.e., $\mu(A) > 0$ implies $Q_U(A) > 0$ for all A in the σ-algebra $\mathcal{B}$), which by the classical result of Blackwell and Dubins (1962) entails that μ-almost surely the variational distance $\sup_{A \in \mathcal{B}} |\mu(A \mid \boldsymbol{x}^n) - Q_U(A \mid \boldsymbol{x}^n)| \to 0$ as $n \to \infty$ (see, e.g., Huttegger, 2015, 617f), so in particular (I: Δ_1). □

6. HD-methods and Bayesian methods

So how does the Solomonoff-Levin function evade Putnam's diagonal argument? As we saw above, the very motivation for the expansion to the class of Σ_1 measures is to evade diagonalization—to obtain universal elements. The Solomonoff-Levin measure is a universal element; as such, it tracks every Δ_1 measure in the sense of (I: Δ_1). The downside is that, as a universal Σ_1 element, the Solomonoff-Levin measure is itself no longer Δ_1 (or the class of Δ_1 measures would already have universal elements).

The force of Putnam's diagonal proof is that no confirmation function can satisfy both (I*) and (II*), and the Solomonoff-Levin proposal is no exception. The Solomonoff-Levin function is powerful enough to avoid diagonalization and fulfill convergence condition (I: Δ_1), but the price to pay is that it might be said to be *too* powerful. It is no longer effective in the sense of (II: Δ_1). Does this invalidate the Solomonoff-Levin function as an inductive rule—let alone a universal one?

One reply is that we cannot hold this against the Solomonoff-Levin definition, since, after all, Putnam has shown that this incomputability is really a *necessary condition* for a policy to be optimal in the sense of convergence condition (I*): "an optimal strategy, if such a strategy should exist, cannot be computable ... any optimal inductive strategy must exhibit recursive undecidability" (Hintikka, 1965, 283). However, this reply seems to miss the second component of Putnam's charge. This is the claim that, while no *confirmation function* can fulfill both adequacy conditions, *other methods* could—in particular, the method HD.

In the current section we consider this claim. As discussed already in some detail by Kelly et al. (1994, 99ff), it actually turns out to be the weak spot in Putnam's argument. When we have this claim out of the way, we can, in the next section, follow up on the above reply and consider the question of Q_U's adequacy afresh.

6.1. **HD methods and confirmation functions.**

Recall that I formulated (I*) and (II*) as conditions on inductive methods in general, not just confirmation functions. Again, Putnam (1963a, 770ff) takes it to be important for his case against Carnap that these conditions are not supposed to be mutually exclusive *a priori*; or it could be seen as a rather moot charge that no confirmation function can satisfy both, either. No confirmation function can satisfy both—conditions (I: Δ_1) and (II: Δ_1) are mutually exclusive—but other methods can: and the method HD that Putnam describes is to be the case in point.

Crucially, however, Putnam's method HD depends on the hypotheses that are actually proposed in the course of time. The method HD fulfills convergence condition (I†), which is so phrased as to accommodate this dependency: the method will come to accept (and forever stick to) any true computable hypothesis, *if* this hypothesis is ever proposed. Thus the method HD relies on some "hypothesis stream" (Kelly et al., 1994, 107) that is external to the method itself; and the method will come to embrace a true hypothesis whenever this hypothesis is part of the hypothesis stream.

In computability-theoretic terminology, the method uses the hypothesis stream as an *oracle*. The HD method is a simple set of rules, so obviously computable—*given* the oracle. But the oracle itself might be incomputable. Indeed, since the computable hypotheses are not effectively enumerable, the hypothesis stream of computable hypotheses *is* incomputable. This is why Putnam must view the oracle as external to the method HD. The alternative is to view the generation of a particular hypotheses stream η as *part of the method itself*; but if any such HD-with-particular-hypothesis-stream-η method—or simply 'HD^η method'—is powerful enough to satisfy convergence condition (I*), then the hypothesis stream and hence the method HD^η as a whole must be incomputable. Putnam is well aware of this: "it is easily seen that any method that shares with Carnap the feature: what one will predict 'next' depends *only* on what has so far been observed, will also share the defect: either what one should predict will not in practice be *computable*, or some law will elude the method altogether" (Putnam, 1963a, 773). The diagonal proof described in section 3 readily applies to any method M: simply construct a computable sequence that goes against M's computable predictions at each point in time (also see Kelly et al., 1994, 102f).

In short, the HD^η methods are in much the same predicament as Carnap's confirmation functions. Conditions (I*) and (II*) *are* mutually exclusive—unless we allow the method to be such that "the acceptance of a hypothesis also depends on *which* hypotheses are actually proposed" (Putnam, 1963a, 773), i.e., allow the method access to an external hypothesis stream.

But Putnam's assumption of an (incomputable) external oracle does, of course, raise questions of its own. The idea would be that we identify the oracle with the elusive process of the invention of hypotheses, the unanalyzable "context of discovery"; ultimately rooted, maybe, in "creative intuition" (Kelly et al., 1994, 108) or something of the sort. Is this process somehow incomputable? How would we know? More importantly, "if Putnam's favourite method is provided access to a powerful oracle, then why are Carnap's methods denied the same privilege?" (ibid., 107).

Kelly et al. offer Putnam the interpretation that the method HD provides an "architecture," a recipe for building particular methods (in my above terminology, HD^η methods), that is "universal" in the sense that for every computable hypothesis, there is a particular computable instantiation of the architecture (a particular computable HD^η method) that will come to accept (and forever stick to) the hypothesis if it is true. "A scientist wedded to a universal architecture is shielded from Putnam's charges of inadequacy, since ... there is nothing one could have done by violating the strictures of the architecture that one could not have done by honoring them" (ibid., 110). Kelly et al. are not convinced, though, that their suggestion saves Putnam's argument, for the reason that it makes little sense for Putnam to endorse a universal architecture while calling every particular instance inadequate and therefore "*ridiculous*" (ibid., 110f; here they quote Putnam, 1974, 238). There is, however, a more fundamental objection. Again, Putnam's argument against Carnap would only be completed if the above way out for the method HD were not open to confirmation functions. That is, it would only succeed if confirmation functions could not be likewise seen as instantiations of some universal architecture. But as a matter of fact, they can. They can be seen as instantiations of the *classical Bayesian* architecture. (I follow Diaconis and Freedman, 1986, 11 in adopting the designation '*classical* Bayesian.' Also see Skyrms, 1996.)

This architecture BAYES employs a countable *hypothesis class* (where hypotheses are again measures over Cantor space), as well as a *prior distribution* that gives positive probability to every element of this hypothesis class. Given a hypothesis class $\mathcal{H}$ and prior w, the corresponding Bayes-with-particular-hypothesis-class-$\mathcal{H}$ method—or 'Bayes$^{\mathcal{H}}$ method'—is given by the measure that is simply the w-weighted mean over the hypotheses in $\mathcal{H}$, i.e., $\xi_w^{\mathcal{H}}(\boldsymbol{x}) := \sum_{h \in \mathcal{H}} w(h)h(\boldsymbol{x})$.

The classical Bayesian architecture is a universal architecture because for every (computable) deterministic hypothesis, there is a particular (computable) instantiation of the architecture (a Bayes$^{\mathcal{H}}$ method where $\mathcal{H}$ contains the hypothesis) that will converge on it when it is true. Just like the HD architecture is guaranteed to converge on (i.e, accept and stick to) every true deterministic hypothesis, *whenever* it is included in the hypothesis stream, so the classical Bayesian architecture is guaranteed to converge on every true deterministic hypothesis, *whenever* it is included in the hypothesis class. And this extends to hypotheses that are themselves probabilistic: a Bayes$^{\mathcal{H}}$ method will come to accept and forever stick to any hypothesis μ with μ-probability 1 whenever it is in $\mathcal{H}$. This property is also known as Bayesian *consistency*. It follows from the exact same argument as the proof of theorem 2, given the fact that $\xi_w^{\mathcal{H}}$ *majorizes* every element in $\mathcal{H}$: for every $h \in \mathcal{H}$ we clearly have for all $\boldsymbol{x} \in \mathbb{B}^*$ that $\xi_w^{\mathcal{H}}(\boldsymbol{x}) \geq w(h)h(\boldsymbol{x})$.

The upshot is that there is an analogy between the situation for the method HD and for the method Bayes. No *particular* confirmation function—Bayes$^{\mathcal{H}}$ method—can satisfy both (I*) and (II*). But, similarly, no *particular* HD$^{\eta}$ method can satisfy both (I*) and (II*). Nevertheless, the HD *architecture* is universal. But, similarly, the Bayes *architecture* is universal.

6.2. **The fixity of methods.** Still, there remains a conspicious disanalogy between the HD and the Bayes approach. This difference is *not* the use of theory per se, even though Putnam took that to be the salient characteristic of the method HD. After all, the Bayes approach uses theory, in the form of the hypothesis class $\mathcal{H}$.

Rather, this difference seems to lie in the use of *new* theory. What is somewhat shrouded in the above analogy between the 'oracles' $\mathcal{S}$ and $\mathcal{H}$ is that the method HD is conceived to operate dynamically, with hypotheses that come to it on the fly (hypotheses that are likely prompted by the actual data!), whereas a Bayes method must do with a class of hypotheses that is fixed from the start. The latter is the well-known Bayesian problem of new theory (see Earman, 1992, 195ff) or the "fixity of the theoretical framework" (Gillies, 2001b): the Bayesian procedure, in its standard form, can only be run after we have fixed the model, and no matter how seriously at odds with the data this model will turn out to be, the procedure does not allow us to take a step back and adjust it.

But for our purposes this is really just an instance of the general fact of the *fixity of an inductive method*. An inductive method is a *fixed* method, a function that for every possible finite data sequence has fixed a prediction. And as highlighted before, any such fixed method falls prey to Putnam's diagonal argument.

This issue is actually quite independent even of the role of new theory. We could modify the method Bayes$^{\mathcal{H}}$ to evaluate its own performance at certain points, and, if called for, derive from the data new hypotheses and insert those in $\mathcal{H}$—but in the end this more complicated procedure again specifies a single fixed inductive method (also see Dawid, 1985a, 1255). Likewise, an algorithm that implements the method HD, *plus* an automated search for and discovery of new hypotheses, in the end again fully specifies a particular algorithm for extrapolating data (cf. Gillies, 2001a). These are all fixed methods, and the relevant difference from Putnam's HD architecture is that the latter is an *architecture*, a method that is not fully specified. (Incidentally, the modified method $\text{Bayes}^{\mathcal{H}}$ could also be seen as instantiating a modified Bayes architecture that *is* capable of incorporating new theory.)

In conclusion, Putnam's argument, purporting to show that confirmation functions have fundamental shortcomings that other methods do not, fails. If there is a shortcoming, it is being a fixed inductive method at all. If Putnam wants to maintain that it is possible for some procedure to satisfy both of his conditions, then this cannot be a fixed procedure. It needs to leave things unspecified, as the HD architecture does, and as the (modified) Bayes architecture does, too. And, again, that what is left unspecified needs to be filled in by something incomputable. Putnam would need to say that the scientific process of coming up with hypotheses is an incomputable process.

What Putnam has shown, at the end of the day, is that we are stuck with a dilemma between two possibilities that both sound dubious: either science is fundamentally unable to discover some computable patterns, or science is itself fundamentally incomputable.

7. A universal inductive rule

We have observed that (I*) and (II*) are mutually exclusive: no fixed method can satisfy both. Let me then follow up on the earlier suggestion to not dismiss the Solomonoff-Levin function Q_U out of hand simply because it does not satisfy the special cases (I: Δ_1) and (II: Δ_1)—that it cannot do the impossible. Instead, let me conclude this investigation with a fresh look at the question: could the Solomonoff-Levin definition be an adequate characterization of a *universal inductive rule*?

One can still, with Putnam, divide this question into two parts. First, in the spirit of (I*), is a Solomonoff-Levin function capable of converging on every reasonable (reasonably effective) hypothesis, if it is true—is it *universal* in this sense? Second, in the spirit of (II*), is a Solomonoff-Levin function itself still a reasonably effective method—a proper inductive *rule*?

To start with the first. Could the Solomonoff-Levin function be called universal in the sense that it is able to track *any* pattern? The best vantage point to address this question is to view it as an instantiation of the classical Bayesian architecture that we saw in the previous section. It turns out that the Solomonoff-Levin functions Q_U correspond to the classical Bayesian methods that employ the class of all Σ_1 hypotheses (see Sterkenburg, 2016). To be exact, the measures Q_U are precisely the $\text{Bayes}^{\mathcal{H}_{\Sigma_1}}$ measures $\xi_w^{\Sigma_1}$ with a semi-computable prior w over the hypothesis class $\mathcal{H}_{\Sigma_1}$ of all Σ_1 measures. (In particular, the choice of universal machine U corresponds to the choice of semi-computable prior w over $\mathcal{H}_{\Sigma_1}$.) That means that theorem 2, the statement that Q_U satisfies convergence condition (I: Δ_1), is simply an instance of Bayesian consistency: the measures $\xi_w^{\Sigma_1}$ over all Σ_1 hypotheses, which includes all Δ_1 hypotheses, will almost surely converge on any Δ_1 hypothesis.

The Solomonoff-Levin proposal can be seen as explicitly aiming at an all-inclusive hypothesis class: and it would be successful in this respect insofar the class of Σ_1 hypotheses (or already the class of Δ_1 hypotheses) is indeed such an all-inclusive, *universal* class of hypotheses. It certainly appears in this spirit that Li and Vitányi (2008), presenting the Solomonoff-Levin measure as a "universal prior distribution," make reference to Hume and claim that the "perfect theory of induction" invented by Solomonoff "may give a rigorous and satisfactory solution to this old problem in philosophy" (ibid., 347).

In his book on the problem of induction, Howson (2000) argues that the choice of prior distribution constitutes our inevitable inductive assumptions (ibid., 88):

> According to Hume's circularity thesis, every inductive argument has a concealed or explicit circularity. In the case of probabilistic arguments ... this would manifest itself on analysis in some sort of prior loading in favour of the sorts of 'resemblance' between past and future we thought desirable. Well, of course, we have seen exactly that: *the prior loading is supplied by the prior probabilities.*

(Also see Romeijn, 2004, 357ff.) From the classical Bayesian perspective, the basic structure of the inductive assumption is given by the elements of the hypothesis class. The hypothesis class embodies the regularities that can be extrapolated, the patterns that should gain higher instance confirmation from supporting instances.

As an aside, one can say that the choice of hypothesis class answers by stipulation Goodman's riddle: supposing for a moment that we actually had justification for extrapolating the pattern from the past—Hume's original problem—then *which* of the many patterns we do we actually extrapolate? (Also recall section 4 above.) Still, working with the same hypothesis class does not preclude that after any finite data sequence two different Solomonoff-Levin functions can give opposed confirmation values for the next datum, depending on the choice of universal machine or particular effective prior over the class. This issue of the remaining subjectivity in the Solomonoff-Levin function and the relation to Goodman's riddle warrants a discussion of its own, that would take me too far from the present concern with the property of convergence to the elements of a general class of hypotheses.

Returning to the original problem of induction, it is important for the observation that Bayesian methods cannot escape Hume's argument that inductive assumptions must be *restrictive*: that it is impossible to have a prior over *everything* that could be true (Howson, 2000, 61ff; Romeijn, 2004, 357ff). From the classical Bayesian perspective, it must be the case that no hypothesis class $\mathcal{H}$ can contain every possible hypothesis, that no $\mathcal{H}$ is fully general. Could $\mathcal{H}_{\Sigma_1}$, then, escape Hume's argument—is $\mathcal{H}_{\Sigma_1}$ fully general?

Rathmanner and Hutter (2011, 1118) write that "according to the Church-Turing thesis, the class of all computable measures includes essentially any conceivable natural environment." Howson (2000, 77), when discussing the claim that only the computable hypotheses represent "genuine discussable hypotheses," demurs:

> it is just not true that we can consider only denumerably many hypotheses ... in the language of ordinary analysis hypothesis spaces of uncountably many elements are dealt with as a matter of course. The fact is that these are all possibilities and they cannot be ignored at the behest of an arbitrary restriction on language.

The issue here is not so much whether or not *we* can conceive of these possibilities or genuinely discuss them (though this will be important below!): the point is rather that *these are all possibilities.* There are uncountably many possible things nature could do, and our restriction to the *computable* possibilities is just that: a restriction of possibilities.

This restriction is definitely not equivalent to the (widely accepted) Church-Turing thesis, that is a thesis about what we can possibly calculate in a purely mechanical fashion. What would be needed is some kind of physical variant of the

Church-Turing thesis, and a "bold" one at that, that not just says that what nature can *calculate* must be Turing-computable but indeed that what nature can *do* must be Turing-computable (also see Sterkenburg, 2016, 477). This is a fertile topic for speculation, but I think that at the end of the day there simply is little justification for promoting the eminently *epistemological* notion of computability to a restriction on what hypotheses could ever be *true*, really a *metaphysical* assumption on the world.

8. An optimal inductive rule

Nevertheless, even if the world might not be constrained by computability, it sounds plausible that *we* necessarily do "view the world through the rose-colored glasses of computable forecasting systems" (Schervish, 1985, 1274). Plausibly, we *are* constrained by computability in our inductive methods.

Consequently, if we interpret the elements of $\mathcal{H}_{\Sigma_1}$ as corresponding to *inductive methods* rather than hypotheses, then $\mathcal{H}_{\Sigma_1}$ might be interpreted as containing *all possible* inductive methods. Wherefore the Solomonoff-Levin function can be reinterpreted as aggregating over the pool of *all possible* inductive methods.

8.1. **Towards an optimal inductive rule.** As explained in more detail in the appendix, the convergence theorem 2 can actually be derived from the following more 'absolute' fact. For any Σ_1 measure ν, there is a constant bound on the surplus *logarithmic loss* (expressing the divergence between the given confirmation values and the symbols that actually obtain) incurred by Q_U relative to this measure ν, on *any* symbol sequence. Thus, if we take the Σ_1 measures as giving all possible inductive rules, then Q_U is a universal inductive rule in the following powerful sense: *it is an inductive rule that compared to* any *other inductive rule will* never *perform much worse.*

To put it another way. A Solomonoff-Levin function might not do well if nature generates—incomputably—adversarial data: but there is a sense in which this is not so interesting. Arguably, *no* inductive method would do well in that case. More interesting is the case when at least *some* inductive method would do well. And in a precise sense, on the proposed interpretation, a Solomonoff-Levin function will do well in such a case: it will do well if *any* inductive method does. I will brand this the *optimality* interpretation: rather than *reliable* (guaranteed with certainty to converge on the true hypothesis), Q_U is *optimal* in the sense that it is guaranteed to converge on successful predictions (and in particular, converge on the true hypothesis) *if any inductive rule does.* The inductive rule Q_U is *vindicated* in the sense of Reichenbach (1933, 421f; 1935, 410ff; 1938, 348ff; see Salmon, 1967, 52ff, 85ff; 1991).

This interpretation is actually more in line with Putnam's demand that the cleverest possible inductive rule should be able to eventually pick up any pattern *that our actual inductive methods would.* It is also more in line with Solomonoff's original aim that given "a very large body of data, the model is *at least as good as any other that may be proposed*" (1964, 5, emphasis mine).

If we accept this, then the Solomonoff-Levin function *is* a universal inductive rule—defying Putnam's lesson that there can be no such thing (see, in particular, the discussion of van Fraassen, 2000, 257ff of a Reichenbachian conception of a universal inductive rule). As we have seen, the crucial move to unlock this possibility after all, hence the crucial precondition to our optimality interpretation, is the expansion to the nondiagonalizable class of Σ_1 elements. It is time to answer the question whether this move is reasonable at all. Analogous to convergence condition (I*) about the identification of all hypotheses with the Δ_1 measures: is it reasonable to identify all possible prediction methods with those corresponding to the Σ_1 measures?

8.2. **Towards a universal pool of inductive methods.** Most importantly, is the class of Σ_1 measures not *too* wide—does a Σ_1 measure that fails to be Δ_1 still constitute a proper inductive rule? In particular, we have returned to the second question at the start of the previous section: does the Solomonoff-Levin function itself constitute a reasonable (reasonably effective) method?

With the Solomonoff-Levin definition, we do embark, in Putnam's words, on the "doubtful project of investigating measure functions which are not effectively computable" (1963a, 778). An incomputable measure function is certainly impractical, or indeed "of no use to anybody" (Putnam, 1963a, 768) in any practical way—but that already goes for any measure function that *is* computable but not in some sense *efficiently* so. The minimal requirement that Putnam was after is computability *in principle*, i.e., given an unlimited amount of space and time. Indeed, under the Church-Turing thesis, computability is just what it *means* to be (in principle) implementable as an explicit method—computability is the minimal requirement to be a method at all. On this view, a Δ_1 measure is a measure that corresponds to a method that (given unlimited resources) for any finite sequence returns the probability that the measure assigns to it. But, likewise, a Σ_1 measure still corresponds to a method that (given unlimited resources) for any finite sequence returns *increasingly accurate approximations* of its probability. So, albeit in a weaker sense, a Σ_1 measure is still connected to some explicit method.

But even if this is so, the property of mere semi-computability is, on a closer inspection, not so easy to make sense of. To illustrate, consider, in a setting of categorical induction (where a method for a given observed data sequence issues a symbol 0 or 1, rather than a probability), the unsuitability of a *partial computable* function (see Kelly et al., 1994, 104). For a given data sequence the function might not be defined, and we either have to be prepared to wait forever (in which case, if the function is indeed not defined there, the induction is put on hold indefinitely), or we wait until at some point we decide to break the spell and just issue a default symbol (in which case we actually use a method that reduces to a *total* computable method, or, if this decision is somehow incomputable, a method that is not computable at all). In all cases, we end up with a total function that is either not universal (because computable) or not computable. Now a semi-computable function is at least defined on all trials, which makes it *look* less problematic: but the situation is still fundamentally the same. For each observed data sequence we can only compute lower approximations to the next symbol's probabilities of unknown accuracy, and we either have to be prepared to wait forever to reach the actual value (and unless the probability values sum to 1, in which case we will reach surety about the value up to any accuracy, the induction freezes forever), or we have to (incomputably?) decide at some point to just go with the current approximation. Again, the actual prediction function is either not universal or not computable.

This is already a serious drawback—it casts serious doubt on the adequacy of the effectiveness condition that the Solomonoff-Levin function still satisfies. As such, it is also an additional problem for the universal reliability interpretation of section 7: *both* conditions are dubious. But for the universal optimality interpretation there is actually another problem that precedes this one, a crucial detail that decisely invalidates this interpretation.

8.3. **Diagonalization strikes again.** This crucial detail is the fact that for the purpose of induction, we are not so much interested in the probabilities issued by the measure functions, but by the *conditional* probabilities that give the corresponding confirmation functions' outputs. But this has repercussions for the level of effectiveness.

This aspect is easy to oversee, because for the Δ_1 measures it makes no difference. If a measure μ is Δ_1, so μ as a function on finite sequences is computable, then (and only then) the two-place function $\mu(\cdot \mid \cdot)$, given by $\mu(x_{n+1} \mid \boldsymbol{x}^n) = \mu(\boldsymbol{x}^{n+1})/\mu(\boldsymbol{x}^n)$, is computable as well. Thus the Δ_1 measures correspond precisely to the Δ_1 conditional measures, or confirmation functions. However, for the Σ_1 measures this *does* make a difference.

Namely, a *fraction* $m(\cdot) = \frac{m_1(\cdot)}{m_2(\cdot)}$ of two functions m_1 and m_2 that are both Σ_1 need not itself be Σ_1. The two approximating functions g_1 and g_2 for m_1 and m_2, respectively, give rise to an approximating function $g(\boldsymbol{x}, s) = \frac{g_1(\boldsymbol{x},s)}{g_2(\boldsymbol{x},s)}$; but this function, while it satisfies $\lim_{s\to\infty} g(\boldsymbol{x}, s) = m(\boldsymbol{x})$, does *not* need to satisfy $g(\boldsymbol{x}, s) \leq g(\boldsymbol{x}, s+1)$ for all $s \in \mathbb{N}$. For such a function we do not even know whether any given approximation is a lower aproximation, and whether the approximation at $s+1$ will be at least as accurate as the one at s. In technical terms, the function m is only *limit-computable* or Δ_2 (see Soare, 2016, 63ff). Thus, a confirmation function corresponding to (i.e., a fraction of terms of) a Σ_1 measure is Δ_2, but need no longer be Σ_1. In particular, the *conditional* Solomonoff-Levin function $Q_U(\cdot \mid \cdot)$ is no longer Σ_1.

As a matter of fact, this follows from Putnam's original diagonalization argument, that shows the incompatibility of the conditions (I) and (II) that I introduced in 3. In particular, recall the statement of Putnam's original effectiveness condition, that in our setting of a binary alphabet reads

(II) For every $\boldsymbol{x}^n$, it must be possible to compute a k such that $C(1, \boldsymbol{x}^n 1^k) > 0.5$.

If $Q_U(\cdot \mid \cdot)$ were Σ_1, then Q_U would also satisfy effectiveness condition (II): for any given $\boldsymbol{x}^n$, by computing lower approximations of $Q_U(x_{n'+1} \mid \boldsymbol{x}^n 1^{n'})$ for increasing $n' > n$ we will effectively discover a k with $Q_U(1 \mid \boldsymbol{x}^n 1^k) > 0.5$. This would mean that Q_U satisfies both (I) and (II), which is shown impossible by the diagonal argument. For completeness, the following proof recounts the details of this diagonalization. (See Putnam, 1963a, 768f, Putnam, 1963b, 299 for the original. A different proof has been given by Leike and Hutter, 2015, 370f, but the current proof has the advantage of being very direct.)

Proposition 3. $Q_U(\cdot \mid \cdot) \notin \Sigma_1$.

Proof. Suppose towards a contradiction that $Q_U(\cdot \mid \cdot)$ is Σ_1, so that (II) holds for Q_U. We can now construct a computable infinite sequence $\boldsymbol{x}^\omega$ as follows. Start calculating $Q_U(0 \mid 0^n)$ from below in dovetailing fashion for increasing $n \in \mathbb{N}$, until an n_0 such that $Q_U(0 \mid 0^{n_0}) > 0.5$ is found (since Q_U satisfies (I) such n_0 must exist). Next, calculate $Q(0 \mid 0^{n_0} 1 0^n)$ for increasing n until an n_1 with

$Q_U(0 \mid 0^{n_0}10^{n_1}) > 0.5$ is found. Continuing like this, we obtain a list $n_0, n_1, n_2, ...$ of lengths; let $\boldsymbol{x}^\omega := 0^{n_0}10^{n_1}10^{n_2}1\dots$. Sequence $\boldsymbol{x}^\omega$ is computable, but for each iteration i we have that $Q_U(1 \mid 0^{n_0}1\dots0^{n_i}) \leq 1 - Q_U(0 \mid 0^{n_0}1\dots0^{n_i}) < 0.5$. Thus by construction the instance confirmation of $\boldsymbol{x}^\omega$ will never remain above 0.5, contradicting (I). □

Now one could try to argue that $Q_U(\cdot \mid \cdot)$ is still Δ_2 or *limit computable*, meaning that it still corresponds to a method that converges to any given finite sequence's probability in the limit (see ibid., 365). But the problem runs deeper. The problem is that we cannot recover the optimality interpretation for conditional measures.

Namely, if we would accept that a Δ_2 confirmation function (i.e., a Δ_2 conditional measure) still counts as a possible inductive method, then we should identify the possible inductive methods with the class of Δ_2 confirmation functions (rather than the original class of confirmation functions with underlying Σ_1 measures). That means that the sought-for optimality would have to be relative to *this* class. But the Solomonoff-Levin predictor is not optimal among the Δ_2 confirmation functions—*no* Δ_2 confirmation function is. This is because the class of Δ_2 *measures*, that precisely induces the class of Δ_2 *confirmation functions*, *is* diagonalizable: just like in the Δ_1 case, one can, for any given Δ_2 measure, construct a Δ_2 sequence that it will never converge on. The easiest way to infer this is to realize that the Δ_2 measures are precisely the Δ_1 measures that have access to the halting set $\emptyset'$ as an oracle: in the diagonal proof we can simply replace all occurances of 'Δ_1' with 'Δ_1 in $\emptyset'$.' In computability-theoretic jargon, the diagonal argument can be *relativized* to $\emptyset'$, thus applying to the Δ_2 measures.

Nor can we take a step back and settle for the class of Σ_1 prediction methods. Once again it follows from Putnam's argument above that there cannot exist universal elements in the class of measures that induce the Σ_1 prediction methods: in the exact same way as above, one can for any given Σ_1 conditional measure construct a Σ_1 conditional measure (namely, a computable sequence) it will never converge on.

All of this easily relativizes to any jump $\emptyset^{(n)}$ of the Halting set, showing that the diagonal argument works for the class of Δ_{n+1} prediction methods and the class of Σ_{n+1} prediction methods, for any $n \in \mathbb{N}$. The strategy for optimality cannot work on any level in the arithmetical hierarchy.

9. Conclusion

Thus we conclude this paper on an unhappy note. We discussed how Putnam's diagonal argument shows that no fixed method whatsoever can satisfy at the same time two conditions to qualify as a universal inductive rule: the one on the ability to detect every effectively computable pattern, the other on the effective computability of the method itself. In light of this impossibility result, we considered as candidate universal inductive rules functions that only satisfy a weaker pair of conditions; specifically, we considered the Solomonoff-Levin definition of a universal measure function. The overarching strategy we identified to bring versions of the two conditions together is to locate a natural class of effective functions that cannot be diagonalized, i.e., that contains universal elements. If one could reasonably identify this class of functions with all possible inductive rules, then the universal elements would be vindicated as universally optimal inductive rules: they constitute inductive rules that are in a strong sense at least as good as any other inductive rule.

In particular, the Solomonoff-Levin measures are constructed as universal elements among the Σ_1 measures—and so, our hope ran, they could qualify as such optimal inductive rules. Unfortunately, however, this hope was tempered, first, by the observation that it is actually hard to make sense of an inductive method that is only computably approximable; subsequently, it was squashed by the observation of a fatal flaw in the strategy for optimality: inductive rules should be identified with two-place confirmation (conditional measure) functions rather than the underlying one-place measure functions. As, it turned out, already follows from Putnam's original proof, this affects their effectiveness properties, which ultimately means that no level in the arithmetical hierarchy yields an undiagonalizable class of inductive rules. Putnam's argument stands.

Appendix

Theorem 2 is in the literature (Li and Vitányi, 2008, 352ff; Hutter, 2003, 2062; Poland and Hutter, 2005, 3781) usually presented as a consequence of (variations of) the following stronger result, first shown by Solomonoff (1978, 426f). Let us introduce as a measure of the divergence between two distributions P_1 and P_2 over $\{0,1\}$ the squared *Hellinger distance*

$$H(P_1, P_2) := \sum_{x \in \{0,1\}} \left(\sqrt{P_1(x)} - \sqrt{P_2(x)} \right)^2 . \tag{1}$$

Then, for every $\mu \in \Delta_1$, the expected infinite sum of divergences between Q_U and μ

$$\mathbf{E}_{X^\omega \sim \mu} \left[\sum_{n=0}^{\infty} H\left(\mu(\cdot \mid X^n), Q_U(\cdot \mid X^n) \right) \right] \tag{2}$$

is bounded by a constant.

To see how (I: Δ_1) follows from this constant bound, suppose that Q_U does not satisfy (I: Δ_1): there is a $\mu \in \Delta_1$ such that with probability $\epsilon > 0$ there is a $\delta > 0$ such that $|\mu(x_{n+1} \mid \boldsymbol{x}^n) - Q_U(x_{n+1} \mid \boldsymbol{x}^n)| > \delta$ infinitely often. But that means that with positive probability the infinite sum of squared Hellinger distances is infinite, and the expectation (2) cannot be bounded by a constant.

The proof of the constant bound on (2) starts with the fact that the distance $H(P_1, P_2)$ is bounded by the *Kullback-Leibler divergence*

$$D(P_1 \parallel P_2) := \mathbf{E}_{X \sim P_1} \left[-\log \frac{P_2(X)}{P_1(X)} \right] . \tag{3}$$

The term $-\log P(\boldsymbol{x})$ expresses the *logarithmic loss* of P on sequence $\boldsymbol{x}$, a standard measure of prediction error; the difference $-\log P_2(\boldsymbol{x}) - (-\log P_1(\boldsymbol{x})) = -\log \frac{P_2(\boldsymbol{x})}{P_1(\boldsymbol{x})}$ expresses the surplus prediction error or *regret* of P_2 relative to P_1 on sequence $\boldsymbol{x}$. Thus the Kullback-Leibler divergence (3) expresses the P_1-expected regret of P_2 relative to P_1.

Using $H(P_1, P_2) \leq D(P_1 \parallel P_2)$ one can work out that (2) is bounded by

$$\mathbf{E}_{X^\omega \sim \mu} \left[\sum_{n=0}^{\infty} -\log \frac{Q_U(X_{n+1} \mid X^n)}{\mu(X_{n+1} \mid X^n)} \right] . \tag{4}$$

Now by the universality of Q_U in the class of Σ_1 measures we know that Q_U majorizes μ: for every finite $\boldsymbol{x}$ there is a constant $c \in [0,1]$ such that $Q_U(\boldsymbol{x}) \geq c \cdot \mu(\boldsymbol{x})$. Indeed we can identify c with $w(\mu)$, where w is the prior over hypothesis class $\mathcal{H}_{\Sigma_1}$ in the classical Bayesian representation $\xi_w^{\Sigma_1}$ of Q_U. This fact allows us to derive that *for every sequence*

$\boldsymbol{x}^m$ *of any length* m

$$\begin{aligned}\sum_{n=0}^{m-1} -\log \frac{Q_U(x_{n+1} \mid \boldsymbol{x}^n)}{\mu(x_{n+1} \mid \boldsymbol{x}^n)} &= -\log \prod_{n=0}^{m-1} \frac{Q_U(x_{n+1} \mid \boldsymbol{x}^n)}{\mu(x_{n+1} \mid \boldsymbol{x}^n)} \\ &= -\log \frac{Q_U(\boldsymbol{x}^m)}{\mu(\boldsymbol{x}^m)} \\ &\leq -\log w(\mu). \qquad (5)\end{aligned}$$

This concludes the proof that (2) is bounded by a constant: since the bound (5) holds for any individual sequence of any length, it also holds for (4) and thus for (2).

The absolute optimality property mentioned in section 8 is just this individual sequence bound (5), which continues to hold for ν that are Σ_1. To reformulate, for any such ν, the sum of surplus prediction errors (regrets) of Q_U relative to ν will *always* (for any sequence $\boldsymbol{x}^m$ of any length m) be bounded by a constant:

$$\sum_{n=0}^{m-1} \left(-\log Q_U(x_{n+1} \mid \boldsymbol{x}^n) - (-\log \nu(x_{n+1} \mid \boldsymbol{x}^n))\right) \leq -\log w(\nu).$$

References

P. Achinstein. Confirmation theory, order, and periodicity. *Philosophy of Science*, 30:17–35, 1963.

D. Blackwell and L. Dubins. Merging of opinion with increasing information. *The Annals of Mathematical Statistics*, 33:882–886, 1962.

R. Carnap. *Logical Foundations of Probability.* The University of Chicago Press, Chicago, IL, 1950.

R. Carnap. Replies and systematic expositions. In Schilpp (1963), pages 859–1013.

R. Carnap. Variety, analogy, and periodicity in inductive logic. *Philosophy of Science*, 30(3): 222–227, 1963b.

A. P. Dawid. Calibration-based empirical probability. *The Annals of Statistics*, 13(4):1251–1274, 1985a.

A. P. Dawid. The impossibility of inductive inference. Comment on Oakes (1985). *Journal of the American Statistical Association*, 80(390):339, 1985b.

P. W. Diaconis and D. A. Freedman. On the consistency of Bayes estimates. *The Annals of Statistics*, 14(1):1–26, 1986.

R. G. Downey and D. R. Hirschfeldt. *Algorithmic Randomness and Complexity*, volume 1 of *Theory and Applications of Computability.* Springer, New York, 2010.

J. Earman. *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory.* A Bradford Book. MIT Press, Cambridge, MA, 1992.

D. A. Gillies. Popper and computer induction. *BioEssays*, 23:859–860, 2001a.

D. A. Gillies. Bayesianism and the fixity of the theoretical framework. In D. Corfield and J. Williamson, editors, *Foundations of Bayesianism*, volume 24 of *Applied Logic Series*, pages 363–379. Springer, 2001b.

N. Goodman. A query on confirmation. *The Journal of Philosophy*, 43(14):383–385, 1946.

N. Goodman. On infirmities of confirmation-theory. *Philosophy and Phenomenological Research*, 8(1):149–151, 1947.

J. Hintikka. Towards a theory of inductive generalization. In Y. Bar-Hillel, editor, *Logic, Methodology and Philosophy of Science. Proceedings of the 1964 International Congress*, Studies in Logic and the Foundations of Mathematics, pages 274–288. North-Holland, Amsterdam, 1965.

C. Howson. *Hume's Problem: Induction and the Justification of Belief.* Oxford University Press, New York, 2000.

S. M. Huttegger. Merging of opinions and probability kinematics. *The Review of Symbolic Logic*, 8(4):611–648, 2015.

M. Hutter. Convergence and loss bounds for Bayesian sequence prediction. *IEEE Transactions on Information Theory*, 49(8):2061–2067, 2003.

M. Hutter. On universal prediction and Bayesian confirmation. *Theoretical Computer Science*, 384(1):33–48, 2007.

K. T. Kelly. Learning theory and epistemology. In I. Niiniluoto, M. Sintonen, and J. Woleński, editors, *Handbook of Epistemology*, pages 183–203. Kluwer, Dordrecht, 2004. Page numbers refer to reprint in H. Arló-Costa, V. F. Hendricks, and J. F. A. K. van Benthem, editors, *Readings in Formal Epistemology*, volume 1 of *Graduate Texts in Philosophy*. Springer, 2016, pages 695-716.

K. T. Kelly, C. F. Juhl, and C. Glymour. Reliability, realism, and relativism. In P. Clark and B. Hale, editors, *Reading Putnam*, pages 98–160. Blackwell, Oxford, 1994.

J. Leike and M. Hutter. On the computability of Solomonoff induction and knowledge-seeking. In K. Chaudhuri, C. Gentile, and S. Zilles, editors, *Algorithmic Learning Theory: Proceedings of the Twenty-Sixth International Conference (ALT 2015)*, volume 9355 of *Lecture Notes in Artificial Intelligence*, pages 364–378. Springer, 2015.

L. A. Levin. Some theorems on the algorithmic approach to probability theory and information theory. *Annals of Pure and Applied Logic*, 162:224–235, 2010. Translation of PhD dissertation, Moscow State University, Russia, 1971.

M. Li and P. M. B. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Texts in Computer Science. Springer, New York, third edition, 2008.

A. Nies. *Computability and Randomness*, volume 51 of *Oxford Logic Guides*. Oxford University Press, 2009.

D. Oakes. Self-calibrating priors do not exist. *Journal of the American Statistical Association*, 80(390):340–341, 1985.

J. Poland and M. Hutter. Asymptotics of discrete MDL for online prediction. *IEEE Transactions on Information Theory*, 51(11):3780–3795, 2005.

H. Putnam. 'Degree of confirmation' and inductive logic. In Schilpp (1963), pages 761–783. Reprinted in Putnam (1975), pages 270–292.

H. Putnam. Probability and confirmation. In *The Voice of America Forum Lectures, Philosophy of Science Series 10*. U.S. Information Agency, Washington, D.C., 1963b. Page numbers refer to reprint in Putnam (1975), pages 293–304.

H. Putnam. The 'corroboration' of theories. In P. A. Schilpp, editor, *The Philosophy of Karl Popper, Book I*, volume 14 of *The Library of Living Philosophers*, pages 221–240. Open Court, LaSalle, IL, 1974. Reprinted in Putnam (1975), pages 250–269.

H. Putnam. *Mathematics, Matter, and Method*, volume 1. Cambridge University Press, 1975.

S. Rathmanner and M. Hutter. A philosophical treatise of universal induction. *Entropy*, 13(6): 1076–1136, 2011.

H. Reichenbach. Die logischen Grundlagen des Wahrscheinlichkeitsbegriffs. *Erkenntnis*, 3:401–425, 1933.

H. Reichenbach. *Wahrscheinlichkeitslehre: eine Untersuchung Über die Logischen und Mathematischen Grundlagen der Wahrscheinlichkeitsrechnung*. Sijthoff, Leiden, 1935.

H. Reichenbach. *Experience and Prediction*. University of Chicago Press, Chicago, IL, 1938.

J. Reimann. Randomness—beyond Lebesgue measure. In S. B. Cooper, H. Geuvers, A. Pillay, and J. Väänänen, editors, *Logic Colloquium 2006*, volume 32 of *Lecture Notes in Logic*, pages 247–279. Association for Symbolic Logic, Chicago, IL, 2009.

J.-W. Romeijn. Hypotheses and inductive predictions. *Synthese*, 141(3):333–364, 2004.

W. C. Salmon. *The Foundations of Scientific Inference*. University of Pittsburgh Press, Pittsburgh, PA, 1967.

W. C. Salmon. Hans Reichenbach's vindication of induction. *Erkenntnis*, 35:99–122, 1991.

M. J. Schervish. Comment on Dawid (1985a). *The Annals of Statistics*, 13(4):1274–1282, 1985.

P. A. Schilpp, editor. *The Philosophy of Rudolf Carnap*, volume 11 of *The Library of Living Philosophers*. Open Court, LaSalle, IL, 1963.

A. K. Shen, V. A. Uspensky, and N. K. Vereshchagin. *Kolmogorov Complexity and Algorithmic Randomness*, volume 220 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2017.

B. Skyrms. Carnapian inductive logic for Markov chains. *Erkenntnis*, 35:439–460, 1991.

B. Skyrms. Carnapian inductive logic and Bayesian statistics. In T. Ferguson, L. Shapley, and J. MacQueen, editors, *Statistics, Probability and Game Theory: Papers in Honor of David Blackwell*, volume 30 of *Lecture Notes - Monograph Series*, pages 321–336. Institute of Mathematical Statistics, 1996.

R. I. Soare. *Turing Computability: Theory and Applications*, volume 4 of *Theory and Applications of Computability*. Springer, New York, 2016.

R. J. Solomonoff. A formal theory of inductive inference. Parts I and II. *Information and Control*, 7:1–22, 224–254, 1964.

R. J. Solomonoff. Complexity-based induction systems: Comparisons and convergence theorems. *IEEE Transactions on Information Theory*, IT-24(4):422–432, 1978.

T. F. Sterkenburg. Solomonoff prediction and Occam's razor. *Philosophy of Science*, 83(4):459–479, 2016.

T. Tao. *An Introduction to Measure Theory*, volume 126 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2011.

A. M. Turing. On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, 2(42):230–265, 1936.

B. C. van Fraassen. *Laws and Symmetry*. Clarendon Press, Oxford, 1989.

B. C. van Fraassen. The false hopes of traditional epistemology. *Philosophy and Phenomenological Research*, 60(2):253–280, 2000.

A. K. Zvonkin and L. A. Levin. The complexity of finite objects and the development of the concepts of information and randomness by means of the theory of algorithms. *Russian Mathematical Surveys*, 26(6):83–124, 1970. Translation of the Russian original in *Uspekhi Matematicheskikh Nauk*, 25(6):85-127, 1970.