Raymond J. Solomonoff 1926-2009

Peter Gács
Boston University

Paul M. B. Vitányi
CWI

December 2, 2010

Ray Solomonoff, the first inventor of some of the fundamental ideas of Algorithmic Information Theory, died in December, 2009. His original ideas helped start the thriving research areas of algorithmic information theory and algorithmic inductive inference. His scientific legacy is enduring and important. He was also a highly original, colorful personality, warmly remembered by everybody whose life he touched. We outline his contributions, placing it into its historical context, and the context of other research in algorithmic information theory.

1 Introduction

Raymond J. Solomonoff died on December 7, 2009, in Cambridge, Massachusetts. He was the first inventor of some of the fundamental ideas of Algorithmic Information Theory, which deals with the shortest effective description length of objects and is commonly designated by the term "Kolmogorov complexity."

In the 1950s Solomonoff was one of the first researchers to introduce probabilistic grammars and the associated languages. He championed probabilistic methods in Artificial Intelligence (AI) when these were unfashionable there, and

treated questions of machine learning early on. But his greatest contribution is the creation of Algorithmic Information Theory.

In November 1960, Solomonoff published the report [14] presenting the basic ideas of Algorithmic Information Theory as a means to overcome serious problems associated with the application of Bayes's rule in statistics. His findings (in particular, the invariance theorem) were mentioned prominently in April 1961 in Minsky's symposium report [8]. (Andrei N. Kolmogorov, the great Russian mathematician, started lecturing on description complexity in Moscow seminars about the same time.)

Solomonoff's objective was to formulate a completely general theory of inductive reasoning that would overcome shortcomings in Carnap's [1]. Following some more technical reports, in a long journal paper in two parts he introduced "Kolmogorov" complexity as an auxiliary concept to obtain a universal a priori probability and proved the invariance theorem that, in various versions, is one of the characteristic elements of Algorithmic Information Theory [16, 17]. The mathematical setting of these ideas is described in some detail below.

Solomonoff's work has led to a novel approach in statistics leading to applicable inference procedures such as the minimal description length principle. Jorma J. Rissanen, credited with the latter, relates that his invention is based on Solomonoff's work with the idea of applying it to classical statistical inference [10, 11].

Since Solomonoff is the first inventor of Algorithmic Information Theory, one can raise the question whether we ought to talk about "Solomonoff complexity". However, the name "Kolmogorov complexity" for shortest effective description length has become well entrenched and is commonly understood. Solomonoff's publications apparently received little attention until Kolmogorov started to refer to them from 1968 onward. Says Kolmogorov, "I came to similar conclusions [as Solomonoff], before becoming aware of Solomonoff's work, in 1963–1964" and "The basic discovery, which I have accomplished independently from and simultaneously with R. Solomonoff, lies in the fact that the theory of algorithms enables us to eliminate this arbitrariness by the determination of a 'complexity' which is almost invariant (the replacement of one method by another leads only to the addition of a bounded term)"

Solomonoff's early papers contain in veiled form suggestions about randomness of finite strings, incomputability of Kolmogorov complexity, computability of approximations to the Kolmogorov complexity, and resource-bounded Kolmogorov complexity.

Kolmogorov's later introduction of complexity was motivated by informa-

tion theory and problems of randomness. Solomonoff introduced algorithmic complexity independently and earlier and for a different reason: inductive reasoning. Universal a priori probability, in the sense of a single prior probability that can be substituted for each actual prior probability in Bayes's rule was invented by Solomonoff with Kolmogorov complexity as a side product, several years before anybody else did.

A third inventor is Gregory J. Chaitin, who formulated a proper definition of Kolmogorov complexity at the end of his paper [2].

For a more formal and more extensive study of most topics treated in this paper, we recommend [7].

2 The inventor

Ray Solomonoff published a scientific autobiography up to 1997 as [23]. He was born on July 25, 1926, in Cleveland, Ohio, in the United States. He studied physics during 1946-1950 at the University of Chicago (he recalls the lectures of E. Fermi). He obtained a Ph.B. (bachelor of philosophy) and a M.Sc. in physics. He was already interested in problems of inductive inference and exchanged viewpoints with the resident philosopher of science at the University of Chicago, Rudolf Carnap, who taught an influential course in probability theory.

From 1951-1958 he held half-time jobs in the electronics industry doing math and physics and designing analog computers.

In 1956, Solomonoff was one of the 10 or so attendees of the Dartmouth Summer Research Conference on Artificial Intelligence, at Dartmouth College in Hanover, New Hampshire, organized by M. Minsky, J. McCarthy and C.E. Shannon, and in fact stayed on to spend the whole summer there. (This meeting gave AI its name.) There Solomonoff wrote a memo on inductive inference.

McCarthy had the idea that given every mathematical problem, it could be brought into the form of "given a machine and a desired output, find an input from which the machine computes that output." Solomonoff suggested that there was a class of problems that was not of that form: "given an initial segment of a sequence, predict its continuation." McCarthy then thought that if one saw a machine producing the initial segment, and then continuing past that point, would one not think that the continuation was a reasonable extrapolation? With that the idea got stuck, and the participants left it at that.

Also in 1956, Ray circulated a manuscript of "An Inductive Inference Machine" at the Dartmouth Summer Research Conference on Artificial Intelligence,

and in 1957 he presented a paper with the same name at the IRE Convention, Section on Information Theory, a forerunner of the IEEE Symposium on Information Theory. This partially used Chomsky's paper [3] read at a Symposium on Information Theory held at MIT in September 1956. "An Inductive Inference Machine" already stressed training sequences and using previous solutions in solving more complex problems. In about 1958 he left his half-time position in industry and joined Zator Company full time, a small research outfit located in some rooms at 140 1/2 Mount Auburn Street, Cambridge, Massachusetts, which had been founded by Calvin Mooers around 1954 for the purpose of developing information retrieval technology. Floating mainly on military funding, Zator Co. was a research front organization, employing Mooers, Solomonoff, Mooers's wife, and a secretary, as well as at various times visitors such as Marvin Minsky. It changed its name to the more martial sounding Rockford Research (Rockford, Illinois, was a place where Mooers had lived) around 1962. In 1968, the US Government reacted to public pressure (related to the Vietnam War) by abolishing defense funding of civil research, and Rockford foundered. Being out of a job, Solomonoff left and founded his own (one-man) company, Oxbridge Research, in Cambridge in 1970, and has been there ever since, apart from spending nine months as research associate at MIT's Artificial Intelligence Laboratory, the academic year 1990-1991 at the University of Saarland, Saarbruecken, Germany, and a more recent sabbatical at IDSIA, Lugano, Switzerland.

It is unusual to find a productive major scientist that is not regularly employed at all. But from all the elder people (not only scientists) we know, Ray Solomonoff was the happiest, the most inquisitive, and the most satisfied. He continued publishing papers right up to his death at 83.

In 1960 Solomonoff published [14], in which he gave an outline of a notion of universal a priori probability and how to use it in inductive reasoning (rather, prediction) according to Bayes's rule. This was sent out to all contractors of the Air Force who were even vaguely interested in this subject. In [16, 17], Solomonoff developed these ideas further and defined the notion of enumeration, a precursor of monotone machines, and a notion of universal a priori probability based on his variant of the universal monotone machine. In this way, it came about that the original incentive to develop a theory of algorithmic information content of individual objects was Solomonoff's invention of a universal a priori probability that can be used as a priori probability in applying Bayes's rule.

Solomonoff's first approach was based on Turing machines with markers that delimit the input. This led to awkward convergence problems with which he tried

to deal in an ad-hoc manner. The young Leonid A. Levin (who in [27] developed his own mathematical framework, which became the source of a beautiful theory of randomness), was told by Kolmogorov about Solmonoff's work. He added a reference to it, but had in fact a hard time digesting the informalities; later though, he came to appreciate the wealth of ideas in [16]. Solomonoff welcomed Levin's new formalism with one exception: it bothered him that the universal a priori probability for prediction is a semimeasure but not a measure (see below). He continued to advocate a normalization operation keeping up a long technical argument with Levin and Solovay.

In 2003 he was the first recipient of the Kolmogorov Award by The Computer Learning Research Center at the Royal Holloway, University of London, where he gave the inaugural Kolmogorov Lecture. Solomonoff was a visiting Professor at the CLRC. A list of his publications (published and unpublished) is at http://world.std.com/∼rjs/pubs.html.

3 The formula

Solomonoff's main contribution is best explained if we start with his inference formula not as he first conceived it, but in the cleaner form as it is known today, based on Levin's definition of apriori probability [27]. Let T be a computing device, say a Turing machine. We assume that it has some, infinitely expandable, internal memory (say, some tapes of the Turing machine). At each step, it may or may not ask for some additional input symbol from the alphabet $\{0,1\}$, and may or may not output some symbol from some finite alphabet Σ. For a finite or infinite binary string p, let $T(p)$ be the (finite or infinite) output sequence emitted while not reading beyond the end of p. Consider the experiment in which the input is an infinite sequence of tosses of an independent unbiased coin. For a finite sequence $x = x_1 \dots x_n$ written in the alphabet Σ, let $M_T(x)$ be the probability that the sequence outputted in this experiment begins with x. More formally, let $T^{-1}(x)$ be the set of all those binary sequences p that the output string $T(p)$ contains x as a prefix, while if p' is a proper prefix of p then $T(p')$ does not output x yet. Then

$$M_T(x) = \sum_{p \in T^{-1}(x)} 2^{-|p|}, \tag{1}$$

where $|p|$ is the length of the binary string p. The quantity $M_T(x)$ can be considered the *algorithmic probability* of the finite sequence x. It depends, of course,

on the choice of machine T, but if T is a universal machine of the type called *optimal* then this dependence is only minor. Indeed, for an optimal machine U, for all machines T there is a finite binary r_T with the property $T(p) = U(r_T p)$ for all p. This implies $M_U(x) \geq 2^{-|r_T|} M_T(x)$ for all x. Let us fix therefore such an optimal machine U and write $M(x) = M_U(x)$. This is (the best-known version of) Solomonoff's *apriori probability*.

Now, Solomonoff's prediction formula can be stated very simply. Given a sequence x of experimental results, the formula

$$\frac{M(xy)}{M(x)} \tag{2}$$

assigns a probability to the event that x will be continued by a sequence (or even just a symbol) y. In what follows we will have opportunity to appreciate the theoretical attractiveness of the formula: its prediction power, and its combination of a number of deep principles. But let us level with the reader: it is incomputable, so it can serve only as an ideal embodiment of some principles guiding practical prediction. (Even the apriori probability $M(x)$ by itself is incomputable, but it is at least approximable by a monotonic sequence from below.)

4 First, informal ideas

Scientific ideas of great originality, when they occur the first time, rarely have the clean, simple form that they acquire later. Nowadays one introduces description complexity ("Kolmogorov" complexity) by a simple definition referring to Turing machines. Then one proceeds to a short proof of the existence of an optimal machine, further to some simple upper and lower bounds relating it to probability and information. This a highly effective, formally impeccable way to introduce an obviously interesting concept.

Inductive inference is a harder, more controversial issue than information and randomness, but this is the problem that Solomonoff started with! In the first papers, it is easy to miss the formal definition of complexity since he uses it only as an auxiliary quantity; but he did prove the machine independence of the length of minimal codes.

The first written report seems to be [14]. It cites only the book [1] of Carnap, whose courses Solomonoff attended. And Carnap may indeed have provided the inspiration for a probability based on pure logical considerations. The technical report form allowed the gradual, informal development of ideas.

The work starts with confining the considerations to one particular formal representation of the general inference problem: predicting the continuations of a finite sequence of characters. Without making any explicit references, it sets out to combine two well-studied principles of inductive inference: Bayesian statistics and the principle that came to be known (with whatever historic justification) as "Occam's Razor". A radical version of this principle says that we should look for a shortest explanation of the experimental results and use this explanation for prediction of future experiments. In the context of prediction, it will be therefore often justified to call descriptions *explanations*.

Here is the second paragraph of the introduction:

> Consider a very long sequence of symbols—e.g., a passage of English text, or a long mathematical derivation. We shall consider such a sequence of symbols to be "simple" and have high a priori probability, if there exists a very brief description of this sequence—using, of course, some sort of stipulated description method. More exactly, if we use only the symbols 0 and 1 to express our description, we will assign the probability 2^{-n} to a sequence of symbols, if its shortest possible binary description contains n digits.

The next paragraph already makes clear that what he will mean by a short "description" of a string x: a program of a general-purpose computer that outputs x.

The combination of these three ingredients: *simplicity, apriori probability, universal computer* turned out to have explosive power, forming the start of a theory that is far from having exhausted its potential now, 50 years later. This was greatly helped by Kolmogorov's independent discovery that related them explicitly to two additional classical concepts of science: *randomness* and *information*.

There is another classical principle of assigning apriori probabilities that has been given a new interpretation by Solomonoff's approach: *Laplace's principle of indifference*. This says that in the absence of any information allowing to prefer one alternative to another, all alternatives should be assigned the same probability. This principle has often been criticized, and it is indeed not easy to delineate its reasonable range of applicability, beyond the cases of obvious symmetry. Now in Solomonoff's theory, Laplace's principle can be seen revived in the following sense: if an outcome has several possible formal descriptions (interpreted by the universal monotonic machine), then *all descriptions of the same length are assigned the same probability*.

The rest of the report [14] has a groping, gradual nature as it is trying to find the appropriate formula for apriori probability based on simplicity of descriptions.

The problems it deals with are quite technical in nature, that is it is (even) less easy to justify the choices made for their solution on a philosophical basis. As a matter of fact, Solomonoff later uses (normalized versions of) (2) instead of the formulas of these early papers. Here are the problems:

1. Machine dependence. This is the objection most successfully handled in the paper.

2. If we assign weight 2^{-n} to binary strings of length n then the sum of the weights of all binary strings is infinite. The problem is dealt with in an ad-hoc manner in the report, by assigning a factor $(1-\epsilon)^k$ to strings of length k. Later papers, in particular Solomonoff's first published paper [16] on the subject, solve it more satisfactorily by using some version of definition (1): on monotone machines, the convergence problem disappears.

3. We should be able to get arbitrary conditional probabilities in our Bayesian inference, but probability based on shortest description leads to probabilities that are powers of two. Formula (2) solves this problem as simply as it solved the previous one, but the first publication [16] did not abandon the ad-hoc approach of the technical report yet either, summing up probabilities for all continuations of a certain length (and taking the limit).

4. There are principles of induction suggesting that not only minimal descriptions (explanations) should be considered. Formula (2) incorporates all descriptions in a natural manner. Again, the ad-hoc approach, extending the sum over all descriptions (weighted as above), still is also offered in [16].

   It remained for later researchers (Kolmogorov, Levin) to discover that—in certain models (though not on monotonic computers) even to within an additive constant—asymptotically, the logarithm of the apriori probability obtained this way is the same as the length of the shortest description. Thus, a rule that bases prediction on shortest explanations is not too different from a rule using the prediction fitting "most" explanations. In terms of the monotone machines, this relation can be stated as follows. For a string x, let $Km(x)$ be the length of the shortest binary string that causes the fixed optimal monotonic machine to output some continuation of x. Then

$$Km(x) - 2\log Km(x) \leq -\log M(x) \leq Km(x). \tag{3}$$

The paper [16] offers yet another definition of apriori probability, based on a combination of all possible computable conditional probabilities. The suggestion is tentative and overly complex, but its idea has been vindicated by Levin's theorem, in [27], showing that the distribution $M(x)$ dominates all other "lower semicomputable semimeasures" on the set of infinite sequences. (Levin did not invent the universal semimeasure $M(x)$ as response to Solomoff's work, but rather as a natural technical framework for treating the properties of complexity and randomness.) Here, the *semimeasure* property requires, for all x, the inequalities $M(x) \geq \sum_{b\in\Sigma} M(xb)$, while $M(\Lambda) \leq 1$ for the empty string Λ. Lower semicomputability requires that $M(x)$ is the limit of an increasing sequence of functions that is computable in a uniform way. A computable measure is certainly also a lower semicomputable semimeasure. The dominance property distinguishes Solomonoff's apriori probability among all lower semicomputable semimeasures. Levin's observation is crucial for all later theorems proved about apriori probability; Solomonoff made important use of it later.

The paper [17] considers some simple applications of the prediction formulas, for the case when the sequence to be predicted is coming from tossing a (possibly biased) coin, and when it is coming from a stochastic context-free grammar. There are some computations, but no rigorous results.

5 The prediction theorem

Solomonoff wrote an important paper [18] that is completely traditional in the sense of having a non-trivial theorem with a proof. The result serves as a justification of the prediction formula (2). What kind of justifications are possible here? Clearly, not all sequences can be predicted succesfully, no matter what method is suggested. The two possibilities are:

(i) Restrict the kind of sources from which the sequences may be coming, to a still sufficiently wide class.

(ii) Show that in an appropriate sense, your method is (nearly) as good as any other method, in some wide class of methods.

There is a wealth of research on inference methods considering a combination of both kinds of restriction simultaneously, showing typically that for example if a sequence is generated by methods restricted to a certain complexity class then a successful prediction method cannot be restricted to the same class.

Solomonoff's theorem restricts consideration to sources $x_1x_2\ldots$ with some computable probability distribution P. Over a finite alphabet Σ, let $P(x)$ denote

the probability of the set of all infinite sequences starting with x, further for a letter b of the alphabet denote $P(b|x) = P(xb)/P(x)$. The theorem says that the formula $M(b|x_1 \ldots x_n)$, gets closer and closer to the conditional probability $P(b|x_1 \ldots x_n)$ as n grows—closer for example in a mean square sense (and then also with P-probability 1). This is better than any classical predictive strategy can do. More explicitly, the value

$$S_n = \sum_{x:|x|=n-1} \sum_{b \in \Sigma} P(x)(M(b|x) - P(b|x))^2$$

is the expected error of the squared probability of the nth prediction if we use the universal M instead of the unknown P. Solomonoff showed $\sum_{n=1}^{\infty} S_n < \infty$. (The bound is essentially the complexity $K(P)$, of P, so it is relatively small for simple distributions P. There is no bound when P is not even computable.) Hence the expected squared error can be said to degrade faster then $1/n$ (provided the expectation is "smooth").

The set of *all* computable distributions is very wide. Consider for example a sequence $x_1 x_2 \ldots$ whose even-numbered binary digits are those of π, while its odd-numbered digits are random. Solomonoff's formula will converge to $1/2$ on the odd-numbered digits. On the even-numbered digits, it will get closer and closer to 1 if b equals the corresponding digit of π, and to 0 if it does not. By now, several alternative theorems, and amplifications on this convergence property have appeared: see for example [7, 5].

The proof relies just on the fact that $M(x)$ dominates all computable measures (even all lower semicomputable semimeasures). It generalizes therefore to any family of measures that has a dominating measure—in particular, to any countable family of measures.

Despite the attractiveness of the formula, its incorporation of such a number of classical principles, and the nice form of the theorem, it is still susceptible to a justified criticism: the formula is in a different category from the sources that it predicts: those sources are computable, while the formula is not ($M(xy)/M(x)$ is the ratio of two lower semicomputable functions). But as mentioned above, the restrictions on the source and on the predictor cannot be expected to be the same, and at least Solomonoff's formula is brimming with philosophical significance.

The topic has spawned an elaborate theory of prediction in both static and reactive unknown environments, based on universal distributions with arbitrary loss bounds (rather than just the logarithmic loss) using extensions and variations of the proof method, inspiring information theorists such as Thomas M. Cover [4]. An example is the book by Marcus Hutter [5]. A related direction on prediction

and Kolmogorov complexity, using various loss bounds, going by the name of "predictive complexity", in a time-limited setting, was introduced by Vladimir G. Vovk (see [26] and later works).

We noted that Solomonoff normalized his universal apriori distributions, in order to turn them into regular probability distributions. These normalizations make the theory less elegant since they take away the lower semicomputability property: however, Solomonoff never gave them up. And there is indeed no strong argument for the semicomputability of $M(x)$ in the context of prediction. In about 1992, Robert M. Solovay proved that every normalization of the universal a priori semimeasure to a measure would change the relative probabilities of extensions by more than a constant (even incomputably large) factor. In a recent paper with a clever and appealing proof, Solomonoff [25] proved that if we predict a computable measure with a the universal a priori semimeasure normalized according to his prescription, then the bad changes a la Solovay happen only with expectation going fast to 0 with growing length of the predicted sequence.

6 Universal search

It was not until 1978, that Ray Solomonoff started to pay attention to the emerging field of computational complexity theory. In that year, Leonid Levin arrived in Boston, and they became friends. Levin had discovered NP problems around 1970, independently from Stephen Cook, and had shown the completeness of a small number of NP problems (independently of Richard Karp). For our present purpose, an NP problem is best viewed as a *search problem*. It is defined with the help of a *verification predicate* $V(x,w)$, where x is the *instance*, w is a potential *witness*, and $V(x,w)$ is true if and only if the witness is accepted. We can assume that $V(x,w)$ is computable in time linear in the size $|x|$ of the instance x (in an appropriate computation model, see later). The problem is to decide for a given instance x whether there is any witness w, and if yes, to find one. As an example, consider the problem of finding a description of length l that computes a given string x within time t on some fixed machine U. Let $x = U^t(p)$ mean that machine U computes x in time t from program p. The instance of the problem could be the string 0^l10^t1x, and the verifier $V(0^l10^t1x,p)$ would just check whether $|p| \leq l$ and $U^t(p) = x$.

Levin's paper [6] announces also a theorem that has no counterpart in the works of Cook and Karp: the existence of an algorithm that finds a witness to an NP-complete problem in time optimal to within a multiplicative constant.

Theoretically, this result is quite interesting: from then on, one could say that the question has not been *how* to solve any NP problem efficiently, only *what* is the complexity of Levin's algorithm. If there is a theorem that it works in time $g(|x|)$, then of course also the problem of whether there is any witness at all becomes decidable in time $g(|x|)$.

Levin's paper gave no proof for this theorem (a proof can be found now, for example, in [7]). There is a natural, approximate idea of the proof, though. What is special about an NP problem is that once a potential witness is guessed, it is always possible to check it efficiently. Therefore it does not harm much (theoretically, that is as long as we are willing to tolerate multiplicative constants) a good solution algorithm $A(x)$ if we mix it with some other ones that just make wild guesses. Let $\rho_1,\rho_2,\ldots$ be any computable sequence of positive numbers with $\sum_i \rho_i \leq 1$. We could list *all* possible algorithms $A_1,A_2,\ldots$, in some order, and run them *simultaneously*, making a step of algorithm A_i in a fraction ρ_i of the time. At any time, if some algorithm A_i proposes a witness we check it. In this way, if any algorithm A_i finds witnesses in time $g(|x|)$ then the universal algorithm finds it in time $\rho_i^{-1} g(|x|)$: this is what is meant by optimality within a multiplicative constant.

In order to actually achieve the multiplicative constant in his theorem, Levin indicated that the machine model U has to be of a "random access" type: more precisely, of a type introduced by Kolmogorov and Uspensky and related to the "pointer machine" of Schönhage. He also introduced a variant of description complexity $Kt(w) = \min_{t,z:U^t(z)=w} |z| + \log t$ in which a penalty of size $\log t$ is built in for the running time t of the program z outputting the sequence w on the universal machine U. A more careful implementation of Levin's algorithm (like the one given later by Solomonoff) tries the candidate witnesses w essentially as ordered by their complexity $Kt(w)$.

Up to now, Levin's optimal algorithm has not received much attention in the computational complexity literature. In its present form, it does not seem practical, since the multiplicative constant ρ_z^{-1} is exponential in the length of the program z. (For time bounds provable in a reasonable sense, Hutter reduced the multiplicative constant to 5, but with a tremendous additive constant [7]. His optimal algorithm depends on the formal system in which the upper bounds are proved.) But Solomonoff appreciated it greatly, since in computing approximations to his apriori probability, this seems still the best that is available. He gave detailed implementations of the optimal search (giving probably the first written proof of Levin's theorem), in its application to computing algorithmic probability [19, 21]. These did not result in new theorems, but then Solomonoff had

always been more interested in practical learning algorithms. In later projects (for example [22]) aimed at practical prediction, he defines as the *conceptual jump size* CJS of the program z the quantity t_z/p_z, where p_z is some approximation to the apriori probability of z, and t_z is its running time. The logarithm of the conceptual jump size and Levin's $Kt(w)$ are clearly related.

7 Training sequences

Solomonoff continued to believe in the existence of a learning algorithm that one should find. He considered the approach used for example in practical speech recognition misguided: the algorithm there may have as many as 2000 tuneable real number parameters. In the 1990s, he started a company to predict stock performance on a scientific basis provided by his theories. Eventually, he dropped the venture claiming that "convergence was not fast enough."

In a number of reports: [13, 15, 20, 22, 9, 24], universal search as described above is only a starting point for an array of approaches, that did not lead to new theorems, but were no less dear to Ray's heart for that. What we called "program" above can alternatively be called a "problem solving technique", or a "concept". This part of Ray's work was central for him; but the authors of the present article are closer to mathematics than to the experimental culture of artificial intelligence, therefore the evaluation poses challenges for them. We hope that the AI community will perform a less superficial review of this part of the oevre than what we can offer here.

Learning proceeds in stages, where each stage includes universal search. The conceptual jump size CJS introduced above (see [9]) continues to play a central role. Now, "probability" is used in the sense of the probability assigned by the best probabilistic model we can find in the available time for the given data. There is also an update process introducing more and more complex concepts. The concepts found useful on one stage are promoted to the status of primitives of a new language for the next stage, allowing to form more complex composite concepts (and goals). They are combined in various ways, assigning preliminarily just product probability to the composite concept. If a composite concept proves applicable with a probability beyond this initial value, it will be turned it into a new building block (with a corresponding larger probability). In this way, one hopes to alleviate the problem of excessively large multiplicative constants of universal search (see [21]).

Ray did not limit inductive inference to a model where a learner is presented

a stream of experimental results. He realized that in practice, a lot of learning happens in a much more controlled situation, where there is a "teacher" (or several). Now, *supervised learning* is a well-studied set of models: in this, a teacher provides answers to some set of questions that the learner can ask. In Solomonoff's model, the teacher also *orders* the questions in increasing conceptual jump size, facilitating thereby the above concept-building process. Already the report [13] sketches a system designed to recognize more and more complex patterns, as it is being fed a sequence of examples of gradually increasing complexity.[1] Ray spent many years working out some examples in which a learning algorithm interacts with a training sequence. The examples were of the type of learning a simple language, mainly the language of arithmetic expressions. By now, there are systems in AI experimenting with learning based on universal optimal search: see Schmidhuber in [12] and other works.

We are not aware of any *theoretical* study that distinguishes the kind of knowledge that the teacher can transmit directly from the one that the student must relearn individually, and for which the teacher can only guide: order problems by complexity, and check the student answers. The teacher may indeed be in conscious possession of a network of concepts and algorithms, along with estimates of their "conceptual jump size", and we should assume that she communicates to the student directly everything she can. (The arithmetic algorithms, Ray's main example, can certainly be fed into a machine without need for learning.) But it appears that in typical realistic learning, the directly, symbolically transferable material is only a very incomplete projection of the mental models that every pupil needs to build for himself.

[1] Marvin Minsky considers that the practical potential of the pattern recognition algorithms in this work of Ray still has not received the attention it deserves.

Bibliography

[1] Rudolf Carnap. *Logical Foundations of Probability*. University of Chicago Press, 1950. 1, 4

[2] Gregory J. Chaitin. On the length of programs for computing binary sequences, II. *Journal of the ACM*, 16:145–159, 1969. 1

[3] Noam Chomsky. Three models for the description of language. *IRE Trans. Inform. Theory*, 2(3):113–124, September 1956. 2

[4] Thomas M. Cover. Universal gambling schemes and the complexity measures of Kolmogorov and Chaitin. In J. K. Skwirzynski, editor, *The Impact of Processing Techniques on Communication*, pages 23–33. Martinus Nijhoff, 1985. Stanford University Statistics Department Technical Report # 12, 1974. 5

[5] Marcus Hutter. *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability*. Springer-Verlag, Berlin, 2005. 5

[6] Leonid A. Levin. Universal sequential search problems. *Problems of Inform. Transm.*, 9(3):255–256, 1973. 6

[7] Ming Li and Paul M. B. Vitányi. *Introduction to Kolmogorov Complexity and its Applications (Third edition)*. Springer Verlag, New York, 2008. 1, 5, 6

[8] Marvin L. Minsky. Problems of formulation for artificial intelligence. In R. E. Bellman, editor, *Proceedings of the Fourteenth Symposium in Applied Mathematics*, pages 35–45, New York, 1962. American Mathematical Society. 1

[9] Wolfgang Paul and Raymond J. Solomonoff. Autonomous theory building systems. In P. Bock, M. Loew, and M. Richter, editors, *Neural Networks and Adaptive Learning*, pages 1–13, Schloss Reisenburg, 1990. 7

[10] Jorma J. Rissanen. A universal prior for integers and estimation by minimal discription length. *Annals of Statistics*, 11(2):416–431, 1983. 1

[11] Jorma J. Rissanen. *Stochastic Complexity in Statistical Inquiry*. World Scientific, London, U.K., 1989. 1

[12] Jürgen Schmidhuber. Optimal ordered problem solver. *Machine Learning*, 54:211–254, 2004. 7

[13] Raymond J. Solomonoff. An inductive inference machine. In *IRE Convention Record, Section on Information Theory*, pages 56–62, New York, 1957. Author's institution: Technical Research Group, New York 3, N.Y. 7

[14] Raymond J. Solomonoff. A preliminary report on a general theory of inductive inference. Technical report, Zator Company, Cambridge, MA, 1960. 1, 2, 4

[15] Raymond J. Solomonoff. Training sequences for mechanized induction. In M. Yovits, editor, *Self-organizing systems*, 1961. 7

[16] Raymond J. Solomonoff. A formal theory of inductive inference I. *Information and Control*, 7:1–22, 1964. 1, 2, 2, 3, 4, 4

[17] Raymond J. Solomonoff. A formal theory of inductive inference II. *Information and Control*, 7:225–254, 1964. 1, 2, 4

[18] Raymond J. Solomonoff. Complexity-based induction systems: Comparisons and convergence theorems. *IEEE Transactions on Information Theory*, IT-24(4):422–432, July 1978. 5

[19] Raymond J. Solomonoff. Optimum sequential search. Technical report, Oxbridge Research, Cambridge, MA, 1984. 6

[20] Raymond J. Solomonoff. Perfect training sequences and the costs of corruption - a progress report on inductive inference research. Technical report, Oxbridge Research, Cambridge, MA, 1984. 7

[21] Raymond J. Solomonoff. The application of algorithmic probability to problems in artificial intelligence. In L. N. Kanal and J. F. Lemmer, editors, *Uncertainty in Artificial Intelligence*, Advances in Cognitive Science, AAAS Selected Symposia, pages 473–491, North-Holland, 1986. Elsevier. 6, 7

[22] Raymond J. Solomonoff. A system for incremental learning based on algorithmic probability. In *Proceedings of the Sixth Israeli Conference on Artificial Intelligence, Computer Vision and Pattern Recognition*, pages 515–527, Tel Aviv, 1989. 6, 7

[23] Raymond J. Solomonoff. The discovery of algorithmic probability. *Journal of Computer System Sciences*, 55(1):73–88, 1997. 2

[24] Raymond J. Solomonoff. Progress in incremental machine learning. Technical Report 03-16, IDSIA, Lugano, Switzerland, 2003. Revision 2.0. Given at NIPS Workshop on Universal Learning Algorithms and Optimal Search, Dec. 14, 2002, Whistler, B.C., Canada. 7

[25] Raymond J. Solomonoff. The probability of "undefined" (non-converging) output in generating the universal probability distribution. *Information Processing Letters*, 106(6):238–246, 2008. 5

[26] Vladimir G. Vovk. Prediction of stochastic sequences. *Problems of Information Transmission*, 25:285–296, 1989. 5

[27] Alexander K. Zvonkin and Leonid 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 Math. Surveys*, 25(6):83–124, 1970. 2, 3, 4