The Foundations of Solomonoff Prediction

MSc Thesis
for the graduate programme in
History and Philosophy of Science
at the
Universiteit Utrecht

by

Tom Florian Sterkenburg

under the supervision of
prof.dr. D.G.B.J. Dieks
(Institute for History and Foundations of Science, Universiteit Utrecht)
and
prof.dr. P.D. Grünwald
(Centrum Wiskunde & Informatica; Universiteit Leiden)

February 2013

Parsifal:

Wer ist der Gral?

Gurnemanz:

Das sagt sich nicht;
doch bist du selbst zu ihm erkoren,
bleibt dir die Kunde unverloren. –
Und sieh! – Mich dünkt,
daß ich dich recht erkannt:
kein Weg führt zu ihm
durch das Land,
und niemand könnte ihn beschreiten,
den er nicht selber möcht' geleiten.

Parsifal:

Ich schreite kaum, -
doch wähn' ich mich schon weit.

Richard Wagner, Parsifal, Act I, Scene I

Voor mum

Abstract

R.J. Solomonoff's theory of Prediction assembles notions from information theory, confirmation theory and computability theory into the specification of a supposedly all-encompassing objective method of prediction. The theory has been the subject of both general neglect and occasional passionate promotion, but of very little serious philosophical reflection. This thesis presents an attempt towards a more balanced philosophical appraisal of Solomonoff's theory.

Following an in-depth treatment of the mathematical framework and its motivation, I shift attention to the proper interpretation of these formal results. A discussion of the theory's possible aims turns into the project of identifying its core principles, and a defence of the primacy of the unifying principle of Universality supports the development of my proposed interpretation of Solomonoff Prediction as the statement, to be read in the context of the philosophical problem of prediction, that in a universal setting, there exist universal predictors.

The universality of the setting is grounded in the central assumption of computability: while this assumption is not uncontroversial as a constraint on the world, I argue that it is hardly a constraint at all if we restrict attention to all possible competing prediction methods. This is supported by a new, more refined convergence result.

Acknowledgements

Many thanks to Dennis and Peter for supervising a thesis that would have been quite unusual to both. If it wasn't for your enthusiastic reception of my initial plan, I wouldn't have pushed it. Thanks for answering many questions and for good discussion.

Thanks also to Paul Vitányi for answering many questions and for good discussion. Thanks to George Barmpalias, Adam Day, Marcus Hutter, Panu Raatikainen, Jan-Willem Romeijn and Theo Kuipers for answering questions and for valuable suggestions. Thanks to fellow students Nick, Fedde and Abram for valuable discussion.

Finally, thanks to my little sister for many things.

Tom Sterkenburg
Amsterdam, February 2013

Contents

Introduction

There are many engaging angles one can take to introduce a topic as profound as the topic of this thesis. (The skeptic – or simply the level-headed? – would perhaps rephrase: many ways of providing engaging context for a topic that is so theoretical as to appear altogether esoteric otherwise... The tone of this thesis will often be the tone of the skeptic, but let me start on a positive note.)

One such perspective stresses the topic's deep roots in the theoretical foundation of computation and information, and consequently its natural connection to the characteristic ideas and developments of the contemporary digital age. The aim of the Alan Turing Year 2012 was not only to commemorate Turing's inception of computer science, as a feat of mathematical logic of the greatest theoretical importance, but also to reflect on the all-pervasive role, little over half a century later, of computers and computation in virtually every aspect of our lives and of society as a whole – including, of course, the practice of science.

Sadly, the writing of this thesis took some time, and in the end it has missed its chance to join in the celebrations of the Turing year...

Another vivid connection to current issues in science is established via the important role of statistical inference. A number of recent controversies, including some notable cases in our own country, have illustrated the need for a continuing look at both the practical application, and, crucially, the theoretical foundations of statistical theory. In fact, in the shadow of the main opposing schools of frequentism and Bayesianism, several less well-known attempts have been undertaken to come to a different and less problematic basis.

An example of such an attempt is J.J. Rissanen's principle of *Minimum Description Length* (MDL). The defining idea is to get rid of the necessity of any probabilistic assumptions on the world by reformulating the aim of statistical inference in terms of data compression.

A main source of inspiration for the development of the MDL principle has been the 1960's theory of prediction by R.J. Solomonoff. This theory presents a mathematical characterization of supposedly universally objective and optimal prediction. The MDL principle is regularly presented as a practical approximation of Solomonoff's idealized theory.

The precise (conceptual) relation between MDL and Solomonoff's theory of Prediction is an issue that itself merits much more attention. An important step in this direction, and more in general an initial step towards an investigation of the foundations of MDL, should be an investigation of the foundations of the theory of Solomonoff Prediction (SP). That investigation is the topic of this thesis.

Some Context and Sources

The following historical sketch is to give a first appreciation of the complex of ideas and efforts that are connected to the theory of Solomonoff. It also serves to introduce some of the main sources I relied on in the writing of this thesis.

Inception of the theory Raymond J. Solomonoff (1926–2009) was driven by the quest for a general method to learn or discover scientific facts, a project that relates to the formation of the field of artificial intelligence (AI) at the time. He belonged to the select group of attendees of the famous Dartmouth Summer Research Project on Artificial Intelligence of 1956, that gave the field its name. Here Solomonoff circulated a report on *An Inductive Inference Machine*, that he worked out into a talk at the 1956 IEEE Symposium on Information Theory, resulting in what could well be the very first scientific publication on a learning machine [161].

Solomonoff properly initiated the theory that I refer to as Solomonoff Prediction in the 1960 paper *A Preliminary Report on a General Theory of Inductive Inference*, and further refined it four years later in Parts I and II of *A Formal Theory of Inductive Inference* [162, 163, 164].

Although Solomonoff's work was mentioned occasionally (for instance by M.L. Minsky in [122, 123, 124], attributing the theory "great philosophical importance" for inductive inference in the latter book), it didn't at first receive too much attention within the AI community, let alone outside it. Not much later than Solomonoff, A.N. Kolmogorov [105], the great Soviet mathematician, and also G.J. Chaitin [16, 17]), independently developed a very much related idea. The independent discovery of this general idea by these three pioneers can be seen to mark the birth of the discipline of *algorithmic information theory* (AIT, also known as *Kolmogorov complexity*).

Rather than inductive inference, Kolmogorov was interested in a characterization of randomness and (later) of information content, and this was the way the ensuing theory of Kolmogorov complexity was to go; but Kolmogorov started to refer to Solomonoff as soon as he became aware of the primacy of his work. Kolmogorov's student L.A. Levin, motivated by the perceived lack of mathematical rigor in Solomonoff's 1960's papers, set out to securely establish the mathematical setting, and further studied this setting for its own sake [191, 115]. In the meantime, Solomonoff proved the important convergence result of his theory, which he only published in 1978 [165].

Chaitin was orginally concerned with program-size complexity, and later also took the basic idea in the direction of measures of randomness and information content. He has done more than anyone to popularize the field of algorithmic information theory via various articles and books (e.g., [18, 20, 22, 21]). However, he has been accused of painting a very idiosyncratic picture, "rewriting the history of the field" and downplaying or ignoring the work of others so as to present "himself as the sole inventor of its main concepts and results" [54]. The grand philosophical interpretations he adjoins to his work, as extending and deepening Gödel's incompleteness results and showing the fundamental randomness of mathematical truth, have been criticized as being highly misleading or simply false [178, 138].

Inspiration in statistics Although Rissanen wasn't aware of the work of Solomonoff at the time, the paper of Kolmogorov was an important influence when he wrote the first ([142], 1978) of the series of papers that instigated the principle of Minimum Description Length. Rissanen's ideas were also shaped by the *Akaike Information Criterion* (AIC, [1], 1973), a method of model selection that is the very first based on information-theoretic ideas. In his later work (e.g., [144]), Rissanen explicitly acknowledges Solomonoff's contributions.

A closely related but independently developed idea is the principle of *Minimum Message Length* (MML, [183], 1968), developed still a decade earlier by C.S. Wallace without knowledge of the concurrent birth of algorithmic information theory.

Kolmogorov, in two talks at the end of his career (1973, [107, 106]), initiated a method of nonprobabilistic model selection based on Kolmogorov complexity that is now known as the *Kolmogorov structure function* approach (see [179, 181]). It has been presented (e.g., [67]) as a refinement of *idealized* MDL, the version of MDL that includes all computable hypotheses and is closest in spirit to Solomonoff Prediction.

The modern theory of MDL was first summarized in [5]. A clear and introductory overview is given in [41], that draws on the much more comprehensive presentation in [66], the first modern textbook on the subject.

Current status of the theory The original theory of Solomonoff again gained some publicity through the textbook *Kolmogorov Complexity and Its Applications* by M. Li and P.M.B. Vitányi (first edition [118] appeared in 1993; third and most recent edition [119] in 2008), that has come to be the standard reference in the area of Kolmogorov complexity. The book devotes a great deal of attention to SP itself, in a presentation that owes much to Levin's mathematical framework. The authors often spice things up with claims about its philosophical implications.

In the last decade, the theory has found a champion in M. Hutter, who took up the project of promoting the theory's wider relevance. He adheres for most part to the modern presentation of Li and Vitányi (also see [111] by his student S. Legg for a more condensed overview of the theory along these lines), enlarging on the philosophical themes that were already present in their work. Next to the contribution of a number of technical results about the theory, Hutter's main feat is the extension of the original theory of passive prediction to what he calls a theory of *Universal Artificial Intelligence* [75, 81, 113, 112, 85] in an active setting that can also accommodate an agent's actions. He couches his writings on both this extension and the original theory in philosophical assertions that, it's safe to say, don't quite attain the same level of rigour as the mathematics, and that are in any case sufficiently bold to make the unguarded philosopher quite uncomfortable. The actively sought-for association with other controversial ideas like the technological *singularity*[1] doesn't help much to take away this unease.

[1] This idea, of a future *intelligence explosion* following artificial intelligence surpassing that of humans (promoted in a near religious context by futurist R. Kurzweil), was in fact also anticipated by Solomonoff in the 1985 text [166].

Motivated by the discrepancy between the grand philosophical claims of the mathematically inclined proponents of the theory (perhaps for most part induced by the desire to popularize it), and the notable lack of interest from the philosophical community for the theory (perhaps due to its mathematically challenging nature), a principal aim of the current thesis is also to bridge this gap by a more careful philosophical appraisal of Solomonoff Prediction. A related subaim, that justifies the sometimes encyclopaedic nature of the thesis, is to both give a feeling for the rich context of Solomonoff's theory and to cleanse it from some commonly assumed affiliations with wildly speculative ideas.

The Plan of This Thesis

The ultimate goal of this work is to present a new way of looking at the theory of Solomonoff. It's my hope that this investigation and the resulting interpretation of SP will lead to a clearer picture of the conceptual relation to associated theories like MDL, and show in what way SP may contribute to the philosophy of prediction in general.

Warming Up To start off, I give in Chapter 0 a quick intuitive overview of both MDL and Solomonoff's theory. I show the conceptual connection between these two theories, as well as the conceptual difference between the standard way of presenting Solomonoff's theory and my interpretation. This Chapter is supposed to give a first impression of the content and bearing of the theory for those who are not familiar with the field and rather stay clear of the technical details.

Solomonoff's Theory of Prediction Chapter 1 gives a detailed account of the formal content of the theory of SP. This includes the *information-theoretic*, *probabilistic* and *effective* background setting (Section 1.1), the definitions of *algorithmic probability* (Section 1.2) and the *universal mixture distribution* (which appear to be equivalent, giving a *universal prior distribution*, Section 1.3), and the results about predicting performance with this *universal predictor* (Section 1.4).

The Principles of Solomonoff Prediction Having set out the formal substance of the theory of SP, in Chapter 2 I turn to the proper way of looking at the theory. This project I put in terms of the search for the *core principles* of the theory. After motivating this connection and introducing the interpretations of SP as a *method*, a *model* and a *theory* of prediction (Section 2.1), I identify and present the three candidate core principles of *Completeness*, *Simplicity* and *Universality* (Section 2.2).

Universality of Solomonoff Prediction The single principle I will ground my interpretation on is the principle of Universality. Chapter 3 is devoted to the universality of the predictor, which covers one half of what I take to be the main result of the theory of SP, that *in a universal setting, there are universal predictors*. In the course of developing this viewpoint, I dismiss the principle of Completeness (Section 3.1),

show how the threat of *subjectivity* can be overcome (Section 3.2), and how the role of Simplicity must be subsidiary to Universality (Section 3.3).

Universality of the Model The defence of my interpretation and the primacy of the principle of Universality is completed in Chapter 4 by a demonstration of the universality of the model of SP. This includes the unrestrictiveness of the information-theoretic (Section 4.1), probabilistic (Section 4.2) and effective (Section 4.3) setting. The last remaining obstacle in the case of effectiveness is overcome by interpreting the elements of the model as competing predictors rather than environments, which is supported by a refined convergence result (Section 4.4).

0 Warming Up

The purpose of this preliminary Chapter is to give a quick and non-technical impression of the conventional way of looking at Minimum Description Length and Solomonoff Prediction. The presentation of the latter also serves as contrast for the view I develop in this thesis.

The MDL Principle

Suppose that we are confronted with a bulk of *data*, and a number of competing *hypotheses* about the data. The challenge that we face is to pick out the hypothesis that explains the data best.

The principle of *Minimum Description Length* (MDL) provides a criterion to decide between the hypotheses. The fundamental intuition that underlies this criterion is that *understanding* or *explaining* the data is to be identified with the ability to *compress* the data. Learning is a matter of finding regularities, and any regularity hands an opportunity to describe the data in a shorter way.

Taking this lead, we should like to pick the hypothesis that compresses the data most. Actually, and more precisely, we should like to choose the hypothesis H such that the description of the data given H *plus* the description of H itself is as concise as possible.

In short, given a class $\mathcal{H}$ of hypotheses and data D, the MDL principle advises us to select the hypothesis

$$\min_{H \in \mathcal{H}} (L(H) + L_H(D)), \tag{0.1}$$

with $L(H)$ the description length of hypothesis H and $L_H(D)$ the description length of data D given H.

The description length of a hypothesis or a piece of data has a natural interpretation as quantifying its *simplicity*. The MDL principle can thus be seen to give a precise guideline towards an optimal trade-off between simplicity and *goodness of fit*. An extremely simple hypothesis is not likely to be any good because it probably won't help us to significantly compress the data. On the other hand, the ad-hoc hypothesis that explicitly records our particular data, although it gives optimal compression of this data, will be no good because it is no less elaborate than the data itself. The best hypothesis will be found somewhere in between these two extremes.

Universal codes Of course, the sense of all this depends on the descriptions used. For the obvious complaint that one could bring against MDL is that it gives a criterion that is really *subjective*: if it's entirely up to our fancy what descriptions we choose

to refer to given hypotheses, there is no basis for preferring the one with a shorter description.

In order to circumvent this trivialization, MDL advocates a particular kind of description methods or *codes*. Namely, codes that for each possible object to be described mimic the *optimal* code for that object, the code that results in its shortest possible description length. Description methods that fit this profile are called *universal codes*, and the core business of the MDL enterprise is the design of such codes.

Solomonoff's Theory of Prediction

Universal codes pertain to a certain kind of objectivity, but are still relative to the class of hypotheses one chooses in advance. In *idealized* MDL, the class is extended to contain *all computable hypotheses*. This move is a defining element of Solomonoff's theory.

Idealized MDL The main idea is that the very shortest description of an object is the shortest *computer program* that generates it. This connects in a direct way to the central role of data compression. The shortest computer program for an object represents the maximally achievable compression of the object.

Following this line, the universal code of idealized MDL is an agreed-upon *computer language*, which we can picture in a very abstract way (as a *universal Turing machine*), or concretely as a familiar general-purpose programming language like Pascal or Python. Such a programming language is a truly universal code because for any other computable code we can come up with a program that implements this code. That means that if an object has a short description via some other code, it automatically has a description via our computer language that is not much longer: the only overhead is the program for the original code.

For the same reason it doesn't matter much which specific programming language we use, as long as it's a general-purpose (*computationally universal* or *Turing-complete*) language. Given any two programming languages, we can write a compiler in the first to execute programs written in the second language. Again, there is only a constant overhead. This points at a specific *invariance* with respect to the choice of computer language, which gives formal basis for treating an object's shortest program in our computer language as the objectively shortest description of the object.

Of course, in practice, particularly in case of few data, such differences do have an effect on our results. Worse, finding the very shortest programs for data objects is utterly infeasible. We can in fact prove that it's *impossible in principle* to compute the shortest program for any given object: there simply can't exist an algorithm that does this. Hence the name *idealized* MDL: a theory of theoretically optimal model selection that realistic versions of MDL can only approximate.

Solomonoff Prediction If MDL is viewed as a more realistic approximation to idealized MDL, then it's only a small step to present MDL as a practice-oriented offspring of Solomonoff's theory. This is because idealized MDL can be described as a model selection version of Solomonoff's theory of Prediction.

Whereas (idealized) MDL is about *model selection* (given a collection of data, identify a hypothesis), SP is about *prediction*: given a collection of data, identify the next data item. Specifically, given a sequence of symbols, predict the next symbol.

Again, the key idea is to set out from a universal computer language that serves as a description method for data objects (in this case, sequences of symbols). The problem of predicting symbols comes down to the proper assignment of *probabilities* to symbol sequences, and these probabilities are derived from the programs written in our computer language. Namely, the shorter the programs that generate a symbol sequence, the greater its probability. That is, the better *compressible* a symbol sequence, the more probable. This reflects an important role for the *simplicity* of prospective sequences.

The formalization of this idea leads to a *universal prediction method* **M**, that gives a probability distribution on symbol sequences. By conditionalizing on the data we have already seen, we can use **M** to derive the probability of each possible next symbol. Importantly, we can rigorously prove that this method of prediction quickly gives optimal predictions, irrespective of the actual data generating distribution – as long as it's computable.

A Different View of Solomonoff Prediction

The previous overview of idealized MDL and SP reflects the dominant way of presenting the subject. Although it provides a coherent and intuitively appealing picture, it's necessarily a bit simplistic – in some instances a bit *too* simplistic.

Relation between the theories To regard practical MDL as just an approximation of ideal MDL, for instance, is to disregard some essential differences (e.g., [66, s. 17.8]). Moreover, there are a number of nontrivial subtleties involved in translating Solomonoff's predictive theory to an idealized MDL principle of model selection (e.g., [119, s. 5.4]. So, when Rissanen notes in the introduction to his 1989 monograph that

> "One main source of inspiration in developing the *MDL* principle has been the theory of *algorithmic* complexity, due to the founding pioneers, Solomonoff, Kolmogorov, and Chaitin",

he is quick to add:

> "As will be apparent, the role of algorithmic complexity theory is inspirational, only, for almost everything about it [...] must be altered to make the ideas practicable." [144, p. 10]

There is yet much to be said about the exact relation between Solomonoff Prediction and the practical and idealized versions of MDL. The discussion in this thesis, however, is restricted to Solomonoff's theory. It also involves a critical evaluation of the above story, resulting in an alternative view that forms a significant departure.

A different view In one sentence, the above standard interpretation says that the content of Solomonoff's theory is the specification – with the help of notions from probability theory, information theory and computability theory – of a perfectly optimal all-purpose method of prediction, that via the essential use of data compression implements (and so justifies) a fundamental preference for simplicity of hypotheses.

In my view, the content of Solomonoff's theory is a statement that pertains to the structure of the philosophical problem of prediction: namely, that notions from probability theory, information theory and computability theory allow us to describe a model of prediction that is universally general (in the sense that it fits any predictive problem), in which we can, moreover, prove the existence of a class of universally reliable prediction methods (with a role for simplicity that is tentative at best). I will defend that this is a less assuming but more *sound* and *useful* way to look at SP.

1 Solomonoff's Theory of Prediction

This first Chapter provides an overview of the formal aspects and technical details of Solomonoff's theory of Prediction (SP). In introducing the various concepts, I roughly take a route from the original presentation of Solomonoff to the modern presentation of Li and Vitányi (and Hutter), while at the same time putting things in a more general perspective that will blend into my own interpretation in the later Chapters.

After putting down the problem setting (Section 1.1), I will present the four different definitions of a universal predictor that Solomonoff introduced in his early 1960's papers. The investigation of these definitions and their properties directs the remainder of the Chapter. The discussion of the first two models mainly serves to introduce in an easy-going fashion some basic concepts of the theory, and is a prelude to the discussion of the central third definition of algorithmic probability (Section 1.2). The link to the modern presentation is fully established in the discussion of the fourth definition of the universal mixture distribution, that also covers the formal equivalence with the previous definition of algorithmic probability (Section 1.3). The Chapter closes with formal convergence and optimality results that are to demonstrate the power of the universal predictor and the value of the theory (Section 1.4).

1.1 The Setting

The framework of Solomonoff's theory is raised from the fusion of three important fields: *information theory* (Subsection 1.1.1), *probability theory* (Subsection 1.1.2) and *computability theory* (Subsection 1.1.3).

1.1.1 Inspiration from Information

I will start by describing the general problem setting of Solomonoff's theory, and proceed by indicating the affinity with information theory.

1.1.1.1 The Problem

The problem that Solomonoff set himself is the following. We are given a finite sequence of symbols from some preordained finite alphabet, and we ask: how would it continue? If we suppose that the given data string is only an *initial segment* (i.e., a finite beginning) of a potentially infinite sequence, how can we successfully uncover (finite parts of) the remainder of the sequence? Based on the structure and patterns in the given string, we have to find a way of extrapolating it, of *predicting* the continuation.

One could figuratively brand the developing binary sequence an ensuing *history* – the complete history being the potentially infinite binary sequence. At any time, we are given the *past*, a finite initial segment of the complete history. Then our objective is to predict the *future* up to some point, a larger finite initial segment of the complete history.

This setting is a prime example of a predictive problem. The tacit assumption is that any other predictive problem can also be put into this form. In that case the problem of prediction is equivalent to the problem of predicting the continuation of binary strings, and results in this setting would extend to prediction in general.

Prediction v. induction Solomonoff's theory is often referred to as Solomonoff *induction* in the literature (e.g., [168, 119]). The act of induction or *inductive inference*, however, has a broader connotation than solely the anticipation of the future (that is, prediction). Importantly, induction is commonly associated with the notion of *explanation*, from a more general hypothesis (see, for instance, [45, s. 4.1]). A relevant, more precise, distinction is that between prediction and *model selection*. Solomonoff's theory is not a theory of model selection, it's a theory of extrapolation of sequences (what is called a *prequential*, predictive-sequential, approach in [38]), a theory of prediction. For that reason, I won't use the label 'Solomonoff Induction', but rather *Solomonoff Prediction* (SP).

1.1.1.2 Information

The previous problem setting can be conceptually simplified by assuming the two-element alphabet $\{0, 1\}$ of *binary digits* or *bits*. Thus Solomonoff's theory is about prediction of *binary sequences* or *bit strings*. This is a harmless simplification because any other finite alphabet can be reduced to the binary alphabet by encoding the original symbols as bit strings.

Notation For future reference, I will briefly go through the bit string notation that I will employ throughout the thesis.

Finite binary strings (formally, elements of set $2^{<\omega}$) I denote alternately by Greek letters σ, τ, ρ. Infinite binary strings (elements of 2^{ω}) are denoted by capital Latin letters S, T. The concatenation of two finite strings σ and τ is simply written $\sigma\tau$.

The length of string σ is marked $|\sigma|$. The n-th bit of σ is denoted $\sigma(n)$. If σ is a *prefix* of τ (that is, $\sigma\rho = \tau$ for some ρ), I write this as $\sigma \preccurlyeq \tau$ (if additionally $|\sigma| < |\tau|$ this becomes $\sigma \prec \tau$). The prefix of σ of length n is denoted $\sigma \restriction_n$. By $\sigma*$ is meant any finite string that has σ as prefix. If σ and τ are *incomparable*, that is, neither is a prefix of the other, this is denoted $\sigma \mid \tau$.

The formulation of the theory in terms of sequences of bits accentuates an intuitive link with the concept of *information*.

Information theory The basis of *information theory* (see [33]) lies in C.E. Shannon's classical paper *A Mathematical Theory of Communication* (1948, [155]). Among other things, Shannon introduced in this pioneering work the *bit* as fundamental unit of information, the notion of

a random variable's (*Shannon*) *entropy* as a measure of information, and the operational characterization of entropy as the minimum expected code length that an encoding of the random variable can achieve. The latter lower bound, established in the *Source Coding Theorem* (or *Noiseless Coding Theorem*), prompted the design of encodings that are close to this optimal expected description length. Prime examples are the (suboptimal) *Shannon-Fano coding* (also proposed in the original paper), and the truly optimal *Huffman coding* [73].

The association of SP with the notion of information is strengthened by the reference to some central elements of information theory. In the paper *The Discovery of Algorithmic Probability* ([169], 1997), that can be seen as his intellectual autobiography, Solomonoff notes that

> "Shannon's papers and subsequent developments in information theory have profoundly influenced my ideas about induction. Most important was the idea that information was something that could be quantified and that the quantity of information was closely related to its probability." [p. 75]

With respect to the quantification of information, recall the heuristic role of data compression in the story of Chapter 0. The actual debt of SP to information-theoretic ideas about encoding will become more clear in Section 1.2. There I will also point out how information theory provides a bridge from binary strings to *probabilities*.

1.1.2 Plugging in Probabilities

The essence of what we may call the *problem of prediction* is that, in principle, *anything can happen*. After patiently witnessing the generation of billions of 0's in a row, we're still not safe to say that the next bit will be a 0 too: who or whatever is responsible might play a trick on us and surprise us with a 1. Nevertheless, it seems only commonsensical to say that it's *likely* that in this situation the next bit will be a 0. Taking this line, Solomonoff's problem setting can be presented more precisely in probabilistic terms: given a finite data sequence, what is the *probability* of each possible continuation?

The problem, then, is to come to a valid probability distribution on bits or (finite) binary strings of them, strings that are continuations of a given data string. We aim for an assignment of probabilities that takes into account the data we already have, and that allows for proper readjustment, in the hope of increasing accuracy, when we learn more. As such, we are led to a probabilistic theory of *confirmation*.

One could say (cf. [135, 60]) that the probabilistic approach to confirmation in the philosophy of science only became prominent with the work of R. Carnap – work that was also an important inspiration to Solomonoff.

1.1.2.1 Carnap and Logical Probabilities

Carnap's project is a high point in the *logical* approach to probability. This approach, already initiated by A. De Morgan and G. Boole, and defended in detail by J.M. Keynes

in his *Treatise on Probability* (1921, [100]), builds on an extension of deductive logic to what is sometimes called *inductive logic*, a logic that includes less-than-certain inferences. The relation of logical implication is generalized to the relation of *confirmation*. As with deductive logic, this requires the framework of a *formal language*.

Carnap's formal language In [8], and later in his monumental *Logical Foundations of Probability* (1950, [10]), Carnap considered very simple logical languages consisting of finitely many monadic predicates (naming properties) and countably many constants and variables (naming individuals). For such a language, he showed how a probability measure $\mathfrak{m}$ over *state descriptions* (maximally consistent sentences that include all combinations of properties and individuals – intuitively: expressing complete information) extends to a probability measure over *all* sentences, and, finally, how this induces a confirmation function

$$\mathfrak{c}(h, e) := \frac{\mathfrak{m}(h \,.\, e)}{\mathfrak{m}(e)} \tag{1.1}$$

that expresses the degree of confirmation of one sentence h (representing some hypothesis) relative to another sentence e (representing some evidence).

This $\mathfrak{c}$ fully depends on the (unconstrained) choice of an initial $\mathfrak{m}$. In pursuance of a unique preferred $\mathfrak{m}$, Carnap further defined *structure descriptions* as those sets of state descriptions that are closed under permutation of names of individuals (intuitively, a different label doesn't reflect any relevant qualitative difference, so these state descriptions can be considered equal). The preferred measure $\mathfrak{m}^*$ then assigns equal measure to each structure description. Carnap's reason for preferring this particular measure, and the induced confirmation function $\mathfrak{c}^*$, is its agreeable property of giving greater measure to "homogeneous" state descriptions that contain few individuals (because within each structure description, the size of which depends on the number of individuals in its elements, the weight is distributed equally).

However, Carnap soon felt the need to generalize his approach. In *The Continuum of Inductive Methods* (1952, [11], also see [13]), the confirmation functions $\mathfrak{c}$ are supplied with a positive real-valued λ (resulting, for each $\mathfrak{c}$, in a *continuum* of functions $\mathfrak{c}_\lambda$) to adjust the extent of the evidence function's updating when given new evidence. Carnap moreover postulates various axioms that enforce aspects like symmetry under permutation of individuals and long-run convergence to relative frequencies.

Interpretation of probabilities Logical theories of probability are often taken to also provide a *meaning* to the concept of probability: the logical *interpretation* of probability (cf. [57, 68]). This places the logical approach in the middle of the peaking skirmish between multiple competing interpretations of probability at the beginning of the previous century.

Various interpretations In the early days of mathematical probability (associated with the names of Bernoulli, Pascal and, principally, Laplace), the concept of probability was customarily interpreted in an *epistemic* sense, as absence of evidence. This *classical* interpretation, inspired by games of chance, mainly applied to calculations that relied on a starting assumption that probability could be equally distributed over all possible outcomes (see the discussion of the *principle of insufficient reason* on page 23).

A main objection to the classical interpretation is its limited applicability. Only in very few circumstances can we confidently determine all possible cases, let alone consider them equally

likely. In practice, we often count the relative *frequency* of events to asses their probability. This link between probability and relative frequency was already explored by Bernoulli and De Moivre in the 18th century, and forcefully defended by J. Venn in the next. A systematic exposition of the *frequency interpretation*, identifying the probability of an event with its limiting relative frequency of occurance, was first given by R. von Mises in the early 20th century; another prominent supporter at the time was H. Reichenbach.

The frequency interpretation is an *objective* interpretation, in the sense that it treats probability as a physical property that can be determined in experiment. Also in the beginning of the previous century, F.P. Ramsey [139] advanced a radical return to an epistemic understanding: the *subjective* (or *personalist*) interpretation. In this view, probabilities are identified with a person's *degrees of belief*. Although he presented his position in strict opposition to Keynes' *objective*-logical interpretation, Ramsey also sought to give a logical account of probability: a logic of beliefs.

Carnap tried to disentangle matters by suggesting that the imprecise everyday *explicandum* "probability" actually denotes two very different concepts, and so gives rise to two different precise *explicata* [9], that we simply have to keep separate.[1] The concept *probability*$_1$ refers to a (logical) measure of confirmation, while the concept *probability*$_2$ refers to a measure of frequency. Naturally, the first concept is the one that is explicated in his own system.

But even after this distinction, Carnap thought, we are still left with the question of the proper meaning of logical probability$_1$. He became dissatisfied with the ambiguity of the interpretation as simply an objective degree of confirmation between a hypothesis and a piece of evidence. Over time, he also abandoned the possible interpretation as an estimation of relative frequency, in favour of an interpretation in terms of *fair betting quotients* [12, p. xv].

Betting The interpretation of probabilities as betting odds was in fact developed as an operational definition of Ramsey's subjective probability. As a generalization of deductive logic's requirement of *consistency*, inductive logic is associated with the requirement of compliance to the axioms of probability[2]. In an inductive logic of beliefs, this can be seen as a minimal requirement of an agent's *rationality*. B. de Finetti [40] first showed that if your degrees of beliefs as betting odds are *coherent* (that is, it's impossible to construct a *Dutch book* out of them, a bet that is guaranteed to lose you money) then they will satisfy the probability axioms. J.G. Kemeny [99] later showed the converse, establishing that coherence is both necessary and sufficient for your degrees of belief satisfying the axioms of probability. This is strong support for the identification of rationality with coherence, and for the operationalization as betting quotients.

Carnap's embracement of the subjectivist's betting odds interpretation makes him part [189, p. 302] of the wider *Bayesian* movement in the philosophy of science, that we turn to now.

[1] Although this distinction was certainly not new (having been noted already by Poisson, and also made explicit by Ramsey), it only became generally acknowledged with Carnap's exposition [189].
[2] Normally, the axioms of Kolmogorov [104].

1.1.2.2 Bayesian Confirmation Theory

The theory of *Bayesian confirmation* (see [47], [34, ch. 5]) in the philosophy of science has two main characteristics. The first is the procedure of conditionalization by the use of Bayes' rule. The second is the adherence to the subjective interpretation of probability.

Bayesian reasoning Consider the following very general form of the problem of scientific inference. We have under investigation a hypothesis H, and we are presented some experimental data D. Now we wonder if (and how much) our confidence in H should increase in light of D. This would depend on

- how much faith we had in H in the first place (the *prior probability* $P(H)$);
- the extent to which H predicted outcome D (denoted $P(D \mid H)$, the probability of D conditional on H, and curiously called the *likelihood of* H);
- how surprising D was (its *expectedness* $P(D)$).

Since we expressed these things in terms of probabilities, we can derive our new confidence in H, the *posterior probability* $P(H \mid D)$, using the probability-theoretic equation known as *Bayes' rule*:

$$P(H \mid D) = \frac{P(D \mid H)P(H)}{P(D)}. \tag{1.2}$$

Having thus derived the posterior probability, we may forget about the old prior probability $P_{old}(H)$ and assign $P_{new}(H) = P_{old}(H \mid D)$. According to the *relevance criterion of evidence*, linking probability to confirmation, the meaning of a strictly greater $P_{new}(H) > P_{old}(H)$ is precisely that H is confirmed by D. Repeating this method of *conditionalization*, as new data is brought in, is the final stage of the procedure of *Bayesian reasoning*.

Bayesian personalism *Bayesianism* or *Bayesian confirmation* reflects a much wider position than the mere acceptance of Bayes' rule, that is simply a consequence of the Kolmogorov axioms and so absolutely uncontroversial. Somewhat confusingly, the term has come to be positively identified with the *subjective* interpretation of probability.

The inclusion of the above method of Bayesian reasoning then comes down to saying that you have to be consistent in updating your beliefs in a hypothesis: your belief in a hypothesis after seeing data should not differ from the conditional belief you had in the hypothesis *before* seeing the data.

Of course, this still leaves you to form your prior (conditional) beliefs: the values to be inserted in the right hand side of Bayes' rule. Unsurprisingly, this is where the difficulties arise.

Problems of personalism The sole requirement of rationality as identified with coherence still leaves the scientist much freedom in constructing a set of degrees of belief. Subjective as they are, they may very well depend on all kinds of objectively irrelevant factors:

> "A given hypothesis might get an extremely low prior probability because the scientist considering it has a hangover, has had a recent fight with his or her lover, is in passionate disagreement with the politics of the scientist who first advanced the hypothesis, harbors deep prejudices against the ethnic group to which the originator of the hypothesis belongs, etc." [148, p. 183]

– a situation that clearly conflicts with the Bayesian apparatus as an objective model of scientific inference. There are, of course, plausibility criteria that can constrain the assignment of a value to the prior probability of a hypothesis further (such as the scientific standing of the person first advancing it, compatibility with accepted theory or theoretical principles like its simplicity), leading to a more *tempered personalism* [156]. W.C. Salmon [147, ch. 7] goes as far as claiming that closing in on the estimated success of an hypothesis by these criteria brings us close to an objective frequentist interpretation again. But procedures for actually deducing values are lacking. There have been, for example, multiple attempts to evaluate the simplicity of a theory in Bayesian terms (e.g., based on the number of independent parameters in the equations of the theory [91], a "clustering assumption" that a homogeneous world is more likely than a heterogeneous one [70], or the identification of simplicity with the coverage of evidence [146]), but none of them is unproblematic [60]. Even Salmon admits that tempered personalism must involve a good bit of handwaving, and that exact numbers for degrees of belief should be taken with a grain of salt.

One reply is that this is a feature rather than a bug. For it is only natural that different scientists start with different and imprecise degrees of belief – the important characteristic of science is that, as evidence accumulates, the different opinions converge to a more and more objective consensus. And indeed, this *washing out* (or *swamping*) *of priors* in the Bayesian framework is supported by a convergence theorem [149] that shows that in the long run degrees of belief of different agents will almost surely merge and converge to the very same hypothesis. This result would seem to imply that the doubtful assignment of subjective priors plays no substantial part in the end, essentially making the problem of priors disappear and providing very strong support for the objectivity of Bayesian inference. As could be expected, things are not as clear-cut as that. First notice the "almost surely" qualification, signifying that the probabilities won't simply merge and converge but have a second-order probability going to 1 of doing so: this is reason for C. Glymour [60] to take the theorem as having relevance only for people who are convinced already of the rationality of Bayesianism. Apart from that, there are some strong and arguably unrealistic conditions assumed by the theorem, such as that the data are taken from independently and identically distributed trials and that all agents have the same likelihoods that are different for different hypotheses [70]. There are more powerful convergence results that don't need the latter conditions, such as the theorem by Gaifman and Snir [56] in a model-theoretic setting based on Doob's theory of *martingales*. But not only are we now confronted with a lack of a useful estimate of the *rate of convergence* [47, ch. 6], a comparison with *formal learning theory* (see page 65) shows that there must exist problems that can't be learned by a Bayesian agent, which implies that the convergence theorem imposes a "dogmatic" *a priori* knowledge in the form of a restriction on the class of possible models [47, ch. 9].

Thus the problem of an assignment of prior probabilities $P(H)$, an assignment that meets the constraints of rationality and objectivity that one might expect from an account of scientific inference, is one of the great challenges for Bayesian confirmation theory.[3] (And then I still leave aside the issues surrounding the expectedness $P(E)$ of data: notably, the problem of *old evidence*.[4])

1.1.2.3 Back to Solomonoff

Transferring the procedure of Bayesian reasoning to Solomonoff's setting of prediction of continuations of binary strings, we may rewrite Bayes' rule (1.2) as

$$P(H_{\sigma\tau} \mid D_\sigma) = \frac{P(D_\sigma \mid H_{\sigma\tau})P(H_{\sigma\tau})}{P(D_\sigma)},$$

where the "hypothesis" $H_{\sigma\tau}$ is just a possible extrapolation $\sigma\tau$ of the past string σ that we have seen so far ("data" D_σ). As the future $\sigma\tau$ implies past σ, we may drop the likelihood term $P(D_\sigma \mid H_{\sigma\tau})$ that must equal 1. Then our formula, estimating the probability of future string τ knowing past σ, reduces to

$$P(\tau \mid \sigma) = \frac{P(\sigma\tau)}{P(\sigma)}. \tag{1.3}$$

The task of assigning probability values appears to be considerably simplified in Solomonoff's setting. Not only can the likelihood $P(\sigma \mid \sigma\tau) = 1$ be crossed out, finding values for the prior probability $P(\sigma\tau)$ and the expectedness $P(\sigma)$ conveniently reduces to the same problem of agreeing on the prior P on finite bit strings. The hope is that the more restricted setting of prediction of bit strings is also susceptible to a solution to the last remaining problem, the problem of the prior probabilities.

Back to Carnap Conditionalization by means of the specific instantiation (1.1) of Bayes' rule lies at the heart of Carnap's system. While symphatetic to the subjectivist interpretation, the formal system and its axioms can be seen as imposing further restrictions on the choice of probability values, thus restraining, in what Carnap thought was a sensible way, the full freedom that the general Bayesian confirmation theory does leave us [189]. Indeed, the logical interpretation can be seen as a sort of idealized extrapolation of the subjective interpretation, where the additional restrictions enforce that the degrees of belief are ultimately determined by the purely logical structure and relations of the propositions in question [174].

[3] The problem of Bayes' original article [6] was to calculate the chance that the probability of an event lies in some interval, given a number of observations. Rather than in the rule that now bears his name (but that doesn't play a direct role in his method), the signifance of Bayes' article is the solution of the problem by calculating a posterior probability from a *prior probability* of distributions. This introduced the fundamental importance of determining the prior distribution – and so this problem *is* an aspect that can truly be called Bayesian. [47, ch. 1]

[4] Values for the likelihoods are often more straightforward to determine, as they generally reflect an objective relation between a hypothesis and experimental data.

It's hard to gainsay, however, that in Carnap's project these restrictions have to be implemented by the use of a highly artificial language, bearing with it an unavoidable element of subjectivity. This is taken by many critics as showing that the endeavour is fundamentally flawed (e.g., [126]).

The project of Solomonoff is in several ways inspired by that of Carnap. In narrowing things down to a prequential setting, Solomonoff follows Carnap's belief that predictive inference is "the most important and fundamental inductive inference" [10, §44]. In his 1964 paper, he emphatically presents his theory as a formalization of Carnap's probability$_1$, thus continuing the tradition of logical probability. Solomonoff was in fact a student of Carnap in Chicago, and he acknowledges Carnap's influence thus:

> "I liked his function that went directly from data to probability distribution without explicitly considering various theories and 'explanations' of the data. (...) I also liked his idea of $P_2(H, D)$ [*this should be* P_1 – *TS*] and the idea of representing the universe by a digital string, but his method of computing the a priori probability distribution seemed unreasonable to me. The distribution depended very much on just what language was used to describe the universe. Furthermore, as one made the describing language larger and more complete, predictions were less and less contingent on the data. Carnap admitted these difficulties, but he felt that his theory nonetheless had redeeming qualities and that we would eventually find a way out of these difficulties.
>
> Algorithmic Probability is close to Carnap's model, and it does overcome the difficulties described." [169, p. 76]

Solomonoff also sought for formal restrictions on probability assignments, relying on a language of bit sequences that would be less vulnerable to charges of arbitrariness.

1.1.3 Connecting to Computability

Solomonoff aimed to arrive at objective prior probabilities by embedding his setting in the theory of *computability*. This move provides a formal grip on the problem, while still keepings things highly general. The central supposition is that any binary sequence is generated in some effectively computable way.

The theory of computability Computability theory[5] originated in the quest for a mathematical characterization of the notion of being computable by an *effective procedure*, an *algorithm*. Intuitively, an effective (or *mechanical*) method is given by a finite number of instructions, that are sufficiently precise to require no ingenuity whatsoever to be followed without error, and that can be executed to produce a result in a finite number of steps.

[5] Also known as *recursion theory* or *recursive function theory*; standard references to the subject are [145, 128, 160].

Various notions This quest was triggered by work in the foundations of mathematics in the first half of the previous century, particularly D. Hilbert's *Entscheidungsproblem*: does there exist an effective procedure to decide for any given mathematical proposition whether it is provable? The first step towards an answer had to consist in making more precise the notion of effectively computable function, and several, seemingly very different, formalizations were advanced. A. Church proposed what was later named *Church's Thesis* [24, 102] to identify the effective functions with his λ-*definable functions* [23, 26], which was rejected by the great logician K. Gödel; Church then proposed to phrase his thesis in terms of Herbrand and Gödel's (*general*) *recursive functions* [69, 61] instead, which was still not to Gödel's liking. Then A.M. Turing [172], independently and unaware of the other efforts, published a model that would convince both.

Turing described an abstract computing device, now known as *Turing machine*, that mimics an idealized human being, only aided by pen and paper, that tirelessly and stupidly carries out a precise set of instructions. (Note that at that time, a *computer* was understood to mean a human computer.) Aimed at the *Enscheidungsproblem*, it was to give the most general description of what can be humanly achieved in such manner, and it did so in a highly figurative way. In the words of Church,

> "(...) computability by a Turing machine (...) has the advantage of making the identification with effectiveness in the ordinary (not explicitly defined) sense evident immediately." [25, p. 43]

S.C. Kleene [101] had already shown that the λ-definable and the general recursive functions identified the same class of functions; now Turing sketched in an appendix of his paper a proof that the model of Turing machines, yielding the class of (*Turing*) *computable functions*, is in the same sense equivalent to the model of the λ-calculus. These results, that widely different formalizations led to the very same class of effective functions, gave strong support to what is now well-known as the *Church-Turing Thesis* [173, 103]: that the effectively calculable functions are precisely the formally computable ones. Says Gödel,

> "(...) with this concept one has for the first time succeeded in giving an absolute notion of an interesting epistemological notion, i.e., one not depending on the formalism chosen." [62, p. 84]

Machines and computers It's safe to say that the Church-Turing Thesis (CTT), in the above form, has found general acceptance [157]. One compelling reason, as noted, is the *confluence* of many formalisms (those of Church, Gödel and Turing only being the the most well-known) that all turned out to capture the precise same class of functions [59]. The lack of counterexamples, more than half a century later, is another. Lastly, the model of Turing can at once be seen as capable of imitating any operation that a human computer following an effective procedure could execute.

Of special significance is Turing's construction of a *universal Turing machine* that can perform any computation by taking as input the code for another Turing machine and some second value x, and emulating the given machine on this x. While the machine model was originally intended to reproduce a human computer at work, nowadays the obvious analogy is the omnipresent digital computer – the sensational rise of which can be seen to have originated in Turing's purely theoretical model. A Turing machine then really corresponds to a single computer program, and a univer-

sal Turing machine to the familiar general-purpose computer that can execute any conceivable program, given the right instructions.

Turing machines (henceforth simply "machines"), and variations thereof, will underly many technical discussions in this thesis. (See page 26 for a more detailed discussion of the model.)

Terminology and notation When I talk about computability, I refer to the formal notion of Turing-computability (i.e., computable by a Turing machine). For the intuitive notion of computability (that may or may not be identified with Turing-computability, depending on whether one adheres to the CTT), I reserve the term *calculability*.

I will often speak about *effectiveness* (e.g., the effectiveness of the setting of SP): this functions as a sort of umbrella term that signifies the role of calculability, which is often made formally precise by Turing-computability or a weaker variant (e.g, *semi-computability*, see page 32).

A Turing machine I usually refer to by the letter M; in case the machine is universal, the letter U. If machine M on input σ halts at some point, I write $M(\sigma)\downarrow$; if it diverges, $M(\sigma)\uparrow$. In the first case, the output is written $M(\sigma)$.

Incomputability With the mathematical formalism in place, it was possible to rigorously prove the existence of fundamentally *undecidable* problems. Church [24] and Turing [172] had directly employed their own respective formalisms to show that the procedure required by the *Entscheidungsproblem* is not effective; accepting the Church-Turing Thesis then settles the problem in the negative. The Turing machine model pointed the way to other *incomputable* problems, problems that no Turing machine could solve, and so, with the CTT, allow for no effective solution. Turing's earliest example was the incomputability of the *Halting problem* [172]: there can be no algorithm that for every combination of a (code for a) machine and another input value can tell whether the machine on this input value ever halts.

We will see that incomputability is also characteristic of Solomonoff's theory.

The mysterious computer At the Dartmouth conference, before he came to the development of his theory, Solomonoff discussed the problem of the extrapolation of sequences with J. McCarthy, one of the founders of the field of artificial intelligence. The day after this discussion, McCarthy made the following observation.

> "Suppose we were wandering about in an old house, and we suddenly opened a door to a room and in that room was a computer that was printing out your sequence. Eventually it came to the end of the sequence and was about to print the next symbol. Wouldn't you bet that it would be correct?" [169, p. 76]

This provides a nice illustration of the heuristic idea that will be formalized in the upcoming Sections. An effective process that produces a bit string up to some point, if allowed to proceed, is likely to give a legitimate continuation of the string. Putting this intuition in more precise terms, the probability of a continuation should express the likelihood that it is generated by some machine that has already produced the given initial segment. Generalizing, the probability of a sequence should express the

likelihood that it is generated by some machine. Since a universal machine emulates any other machine, this reduces to the likelihood that the sequence is produced by some agreed-upon universal machine. This is the main idea that underlies the different definitions that Solomonoff proposes in Parts I and II of his 1964 paper *A Formal Theory of Inductive Inference* [163, 164].

1.2 Algorithmic Probability

This Section introduces three of Solomonoff's early definitions of a prior distribution on bit strings that are all based on descriptions to a universal machine. Essentially, the prior probability of a string is the probability that the universal machine receives an instruction to produce it.

The first definition (Subsection 1.2.1), that specifies a distribution of strings as the relative frequency of their descriptions for a universal machine, forms a direct but still quite imprecize implementation of this idea. The second definition (Subsection 1.2.2) is an attempt to obtain the same distribution via an imaginary experiment of presenting random input to the universal machine, but suffers from a number of important defects. The third definition (Subsection 1.2.3) overcomes the shortcomings of the second definition by the use of a subtly different Turing machine model: this is the important definition of *algorithmic probability*.

1.2.1 A First Definition

The basic idea that concluded the previous Section is that the probability of a binary string should be based on the likelihood that it is produced by a universal machine. The further idea is to quantify this by the likelihood that the universal machine receives a *description* of the string.

1.2.1.1 Ratio of Descriptions

Let's suppose we agree on the particular universal machine to use: call it the *reference machine*. Now we name a string τ a (U)-*description* of another string σ if τ leads U to produce (an extension of) σ: so $U(\tau)\downarrow\succcurlyeq\sigma$. Since the probability of a string is to correspond to the probability of the reference machine processing a description of it, the next step is to give a convincing definition of the latter probability.

One approach is to try to directly capture the relative number of U-descriptions of σ. What is the ratio of descriptions of σ to the total number of descriptions? Here we have to keep in mind that some inputs may lead the machine to get stuck in infinite loops: those inputs shouldn't be counted as descriptions of anything. All other inputs are called *valid*. Let $T_n = \{\tau \in 2^n \mid U(\tau)\downarrow\}$ contain all such valid inputs of length n, and let $T_{n,\sigma} = \{\tau \in T_n \mid U(\tau) \succcurlyeq \sigma\}$ be the subset of inputs that are actually descriptions of σ. The ratio in the limit gives the first definition of the prior probability of σ:

$$P_{\mathrm{I}}(\sigma) = \lim_{n\to\infty} \frac{|T_{\sigma,n}|}{|T_n|} \tag{1.4}$$

Notice that this definition relies on the inconspicuous presupposition that *all descriptions of the same length should be counted as equals.*

1.2.1.2 Indifference About Descriptions

Definition (1.4) already hints at what will be stipulated even more explicitly in the upcoming definitions of Subsections 1.2.2–1.2.3: that all descriptions of the same length have equal probability. Solomonoff notes the resemblance of this assumption to the classical *principle of insufficient reason*:

> "The formulation of the induction system as a universal machine with input strings of fixed length has an interesting interpretation in terms of "the principle of insufficient reason". If we consider the input sequence to be the "cause" of the observed output sequence, and we consider all input sequences of a given length to be equiprobable (since we have no a priori reason to prefer one rather than the other) then we obtain the present model of induction." [163, p. 19]

Insufficient reason and indifference The cornerstone of the classical interpretation of probability (page 14) is that in the absence of any evidence (or in case of symmetrically balanced evidence) probability should be equally distributed over all possible outcomes. This is codified in Laplace's *principle of insufficient reason*: if there is no reason to think differently than that all cases are equally possible, then their probabilities are equal.

Controversial from the start, by the beginning of the previous century the principle was generally considered to have succumbed under the many objections raised [175]. I already mentioned its doubtful applicability in situations not showing clear symmetries like those in games of chance. Then there are the conceptual difficulties concerning the strange move of inferring knowledge from lack thereof ("*Ex nihilo nihil.* It cannot be that because we are ignorant of the matter we know something about it." [48]) and the general concern that even if the reliance on "possibility" is not circular ("the only sensible meaning one can give to the phrase 'equally possible' is, in fact, 'equally probable"' [141, p. 339]) then at least the principle doesn't really say much. Finally, the *Bertrand paradox* shows that in uncountable spaces the principle can be used in infinitely many incompatible ways.

Keynes, while also very critical, attempted to save a version of the principle of insufficient reason that he called the *principle of indifference* [100]. According to [57], this name was to stress the role of an individual judgement of *irrelevance*, a deliberate neglect of pieces of information deemed unimportant to the situation at hand; this is an act of knowledge rather than ignorance.

In the orderly context of bit string descriptions the subtle differences in interpretation of the principles of indifference and insufficient reason are, of course, quite irrelevant: Solomonoff and later Li and Vitányi use both denotations interchangeably. Li and Vitányi proceed to locate a link with another principle, "based on the same intuition" [117, p. 345], namely the *principle of multiple explanations* of the ancient Greek philosopher Epicurus. I will come to pause upon the merit of these associations in the upcoming Chapters.

1.2.2 A Second Definition

The definition of the 1964 paper that I consider next is actually the result of a stepwise refinement of a much simpler model that Solomonoff described first in his earlier 1960 account *A Preliminary Report on a General Theory of Inductive Inference* [162]. I will trace the same route here.

1.2.2.1 Random Descriptions

Imagine that you've taken possession of the earlier reference universal Turing machine U, and you decide to see what happens when you feed it completely random noise. As a way of achieving this, you might use a (truly fair) coin to decide whether to give the machine a 0 (say on throwing heads) or a 1 (tails) next.

Solomonoff first defined a string τ to be a (U)-description of another string σ if τ makes U produce σ and no more: so $U(\tau) \downarrow = \sigma$. Now the probability that in your experiment of coin flipping you arrive at a specific binary string τ to present to U only depends on its length $n = |\tau|$, and is exactly 2^{-n}. So the probability that you arrive at the *minimal*, shortest, *description* $\tau_{\min}$ of some specific σ is given by

$$P'_{\text{II}}(\sigma) := 2^{-|\tau_{\min}|}. \tag{1.5}$$

This probability of arriving in a random way at the minimal U-description for σ is what Solomonoff, initially, took to be the prior probability of σ.

Kolmogorov complexity The notion of a string's shortest description via a universal machine is the essence of the definition of the *Kolmogorov complexity* of an object. That makes Solomonoff one of the inventors of the concept, along with Kolmogorov and Chaitin, who both independently introduced it a few years later. In terms of priority, we may consider Solomonoff the founding father of the field of algorithmic information theory.

The idea behind Kolmogorov complexity as a measure of simplicity is that the complexity of a string is a matter of how easily it can be described. Simple strings that have a clear structure or follow a specific pattern can be given a succinct description in terms of that regularity, while there is probably no better way of describing a completely random string, that follows no rule at all, than just putting down the complete thing. Another way of phrasing this intuition is that simple strings are highly *compressible*: they must show certain regularities that allow us to provide much shorter descriptions.

How to establish in a general way that a given string has some particularly simple pattern? What kind of regularity to look for? In a kind of move that will prove to be thematic to this area, the theory of computability is brought in to fill the holes that the previous informal discussion still leaves open. To illustrate, take some machine M representing a particular decompresssion algorithm. If on input string τ this machine outputs string σ, we say that τ is an M-description of σ: ideally M on τ produces a much longer string σ by following procedures coded in τ or in M itself. A measure of the information content of σ, although relative to M, is the length of its *shortest* such description: this is its *Kolmogorov complexity* (relative to M)

$$C_M(\sigma) = \min\{|\tau| : M(\tau) \downarrow = \sigma\}.$$

This definition of complexity looks arbitrary, because the choice of M is. But every machine can be seen as following some decompression rule that exploits a specific kind of regularity. Most machines, of course, are perfectly inadequate when viewed as implementing compression algorithms, but if the string at hand has any computable regularity, some machine must deal with just this regularity and will do very well on our string. So our strategy is that we replace the question whether a given string σ follows *any* pattern, which is too general to make precise sense of, by the question if there *exists an algorithm that can find regularities* in σ: we look for *calculable* regularity. The final step is to take a universal machine U that by emulating any other machine can be seen as implementing every single algorithm and in particular any decompression procedure, and define the Kolmogorov complexity of our string σ as the length of the shortest universal description for it:

$$C(\sigma) = \min\{|\tau| : U(\tau) \downarrow = \sigma\}.$$

Actually, not just any universal machine U is quite suitable. We not only require that U can emulate any other machine, we also suppose that it does this in a way that is not too clumsy: it should be able to emulate any particular machine with no more than constant overhead. A standard universal Turing machine that emulates M on input σ by taking a description $\langle \rho_M, \sigma \rangle$ for machine M and input σ will do fine (the overhead is represented by the one description for M); but, for instance, the lengths of the descriptions for a universal machine that is defined only for emulating M on σ after receiving double inputs $\langle \rho_M, \sigma\sigma \rangle$ will come to dwarf the lengths of the descriptions for original M. The technical term is that (the function computed by) U should be *additively optimal* – when talking about universal machines in the rest of this thesis, I mean to refer to additively optimal universal machines.

Now an easy consequence of the definition using an additively optimal reference machine is the *Invariance Theorem* that says that, up to a constant, the shortest description given by the reference machine is no longer than that of any other machine: for all M there is a c_M such that for all strings σ we have $C(\sigma) \leq C_M(\sigma) + c_M$. Naturally, this constant c_M represents the aforementioned overhead. The Invariance Theorem supports the claim that C gives an *objective* notion of complexity of strings. The conspicuous drawback of the definition, however, is that the Kolmogorov complexity of an object is incomputable, essentially because the Halting problem is.

Refinements There are at least two problems with the overly simplistic definition (1.5). For a start, there is no obvious reason why only the minimal descriptions should count. When flipping a coin to determine input bits for U, one may not arrive at the shortest description of string σ, but still end up with U-output σ via another, longer, description. Related to this, it makes sense, as we are concerned with prediction of continuations, to again extend the definition of description to also include input strings that result in outputs that start with the desired string. Then τ is a description of σ if $U(\tau)$ converges to any extension $\rho \succcurlyeq \sigma$ of σ.

Solomonoff introduces the sets $T_{\sigma,n}$ for all the descriptions of extensions of length n of σ, and then sums their lengths as n goes to infinity:

$$P''_{\mathrm{II}}(\sigma) := \lim_{n\to\infty} \sum_{\tau_i \in T_{\sigma,n}} 2^{-|\tau_i|} \tag{1.6}$$

The second main problem is that the "distribution" given by (1.5) is not well-defined: the sum of the probabilities doesn't converge. Worse, the new expression (1.6) is

already divergent for a single σ. In order to repair this defect, Solomonoff included a compensating ϵ-term, which finally results in the rather unwieldy definition

$$P_{\mathrm{II}}(\sigma) := \lim_{\epsilon \to 0} \lim_{n \to \infty} \sum_{\tau_i \in T_{\sigma,n}} \left(\frac{1-\epsilon}{2} \right)^{|\tau_i|}. \tag{1.7}$$

1.2.2.2 The Kraft Inequality and Prefix Machines

The fact that expression (1.5) doesn't induce a proper distribution is a direct consequence of the *Kraft inequality*, a central result in information theory. It concerns sets of finite strings that are *prefix-free*. String σ is sometimes called a *prefix* of τ if it is a strict initial segment $\sigma \prec \sigma\rho = \tau$ of τ; a set X of strings is prefix-free if X doesn't contain any prefixes of its elements: $\sigma \in X$ implies that $\sigma \restriction n \notin X$ for all $n < |\sigma|$.[6] Then we have that

Theorem 1.1 (Kraft Inequality [108])**.** *Set $X \subseteq 2^{<\omega}$ of finite strings is prefix-free if and only if*

$$\sum_{\sigma \in X} 2^{-|\sigma|} \leq 1.$$

For a standard universal Turing machine, there is no reason to expect that the set of minimum descriptions of all strings is prefix-free, hence, by Kraft's Inequality, the sum of the "probabilities" given by expression (1.5) will exceed 1.

Prefix machines A straightforward remedy would be to somehow enforce that the minimal descriptions constitute a prefix-free set. It appears that it's not hard to transform any given Turing machine into a machine the complete *domain* of which (i.e., the set of inputs on which it halts, including all minimal descriptions) is prefix-free (see [119, s. 3.1]). In fact, a very natural adaptation of the standard Turing model results in the class of so-called *prefix machines*.

The Turing machine model I will briefly describe the standard Turing machine model, or at least a standard version (models may differ in details that are non-essential, in the sense that the resulting class of machines corresponds to the same class of functions).

The "hardware" of a Turing machine consists of a two-way infinite one-dimensional *tape*, that is divided into *cells* that each contain one of the possible symbols 0, 1 and the *blank* B, and a *head* that moves along the tape and can scan and overwrite the symbols in the cells. The head is always in one of a number of *states*. It is governed by the "software" of the machine, a list of instructions that at each discrete *step* tells the head to either move one cell to the left or the right or to replace the symbol in the current cell with symbol 0, 1 or B, and then to assume a certain state. This instruction uniquely depends on the combination of the state the head is currently in and the symbol it reads from the current cell. The fact

[6] Given any set of strings A, I will denote by $A' = \{\sigma \in A \mid \forall \tau \in A\ \tau \not\succ \sigma\}$ the prefix-free subset obtained by taking out all extensions of other strings in the set. Such a set A' is sometimes called the *bottom* set of A.

that the instruction in each situation is unique makes the machine *deterministic*; however, not every situation has to be anticipated. If the head finds itself faced with a state-symbol combination that doesn't match any instruction, the machine *halts*. The *input* to the machine (the *program* it's given) is the number of cells filled with 1's at the start; we may assume that at that point the head (in a state q_o) is positioned at the leftmost 1-cell (a B-cell if there is none). The *output* of the computation is the number of 1's that are on the tape when the machine has halted.

A prefix-free Turing machine departs from the standard architecture in a number of ways:

- there are three different tapes: a one-way infinite *input tape*, a one-way infinite *output tape*, and a two-way infinite *work tape*;
- associated with the different tapes are three different heads, a read-only *reading head* and a write-only *writing head* belonging to the first and second tape, respectively; and the reading head only allowed to move to the right;
- there is no blank symbol B.

Depending on the symbols scanned by the reading head and the work tape head, and the state the machine is in, the instructions tell the reading head to either move to the right or remain stationary, the work tape and writing head to move or write a symbol, and finally the machine to assume a certain state. Since there are no blanks, the one-way infinite input tape only contains 0's and 1's at the start of the computation; the reading head will then be positioned at the first, leftmost, cell. The *output halts on input* σ when the reading head is at the last symbol of σ when the machine halts.

A moment's reflection reveals that machines of this kind, given any infinite binary string S on the input tape, will only scan and process one maximal initial segment or prefix $\sigma \prec S$. This results in a prefix-free domain $\mathrm{dom}(M) = \{\sigma \mid M(\sigma)\downarrow\}$ for each such *prefix machine* M. An intuitive interpretation is that such machines process *self-delimiting* programs: it's no longer possible to use blanks to signal the end of input, so the program has to somehow encode itself when the machine should stop reading. For that reason we may also call them *self-delimiting machines*.

For the class of prefix machines, too, it is possible to define a universal machine, a *universal prefix machine*. If such a machine is selected as the reference machine, Kraft's inequality ensures that the probabilities given by (1.5), and indeed those given by (1.7), are bounded by 1.

Prefix Kolmogorov complexity What I described in the section about Kolmogorov complexity on page 24, based on regular machines, is commonly called *plain* Kolmogorov complexity. This definition still gives rise to some counterintuitive properties, notably the lack of *subadditivity* (we would like $C(\sigma,\tau) \leq C(\sigma) + C(\tau)$ to hold true), and *complexity dips* in infinite sequences (which imply the impossibility of sequences that only have incompressible initial segments – preventing a natural definition of *random* sequences). These issues can be seen as having to do with the fact that it can't separate the information content of the bits of a bit string from the additional (but accidental) information in the *number* of bits. A regular machine can first

determine the length of the string before actually looking at the contents of the bits, and as such, it can "cheat" and get more information from the string than we intended in defining Kolmogorov complexity (also see [177, s. 5.2.1]).

This is remedied by the use of prefix machines, which, after all, take self-delimiting programs that have to include the information of their length in the bits themselves. This was the motivation for Chaitin [19] to introduce the model (also independently developed by Levin [115]). With the help of a universal prefix machine U, the *prefix Kolmogorov complexity* of a string σ is defined as

$$K(\sigma) = \min\{|\tau| : U(\tau) \downarrow = \sigma\}.$$

1.2.3 The Third Definition: Algorithmic Probability

Solomonoff conjectured [163, p. 18] that the previous two definitions would in fact yield equivalent distributions. However, with hindsight he admits that his definition of the reference machine U is not sufficiently precise to settle this – indeed, he recognizes the room to define operations in such a way that he would be "unable to tell whether the expressions will converge or not" [169, A.2]. For the third model, he defined a very different type of universal machine (that can be used in definition (1.4) as well), a *monotone* machine that can process infinite input sequences.

1.2.3.1 Monotone Machines

The monotone machine can be seen as a dynamic variant of the standard Turing machine. Apart from some work tapes, it has a read-only input and a write-only output tape, both of which can only be processed in one direction. That means that there is no looking back or overwriting: the reading head scans bit after bit and the output tape writes one bit after the other, performing computations on the work tapes inbetween. In the course of reading a (conceivably infinite) input string, the machine spits out output bits that it cannot retract, possibly leading to an infinite output string if the machine never halts.

If monotone machine M's output after processing input τ is σ, and no more (because it halts or computes forever without producing any more output), then this is written $M(\tau) = \sigma$. Monotone machines derive their name from the characteristic property that $M(\tau) \preccurlyeq M(\tau\rho)$ for all sequences τ, ρ.

Again one can construct a universal monotone machine that simulates every single monotone machine; let's again assume some agreed-upon reference universal machine U.

1.2.3.2 Algorithmic Probability

Suppose that you perform the same experiment of generating bits at random and feeding them to a universal machine, but now to a *monotone* machine U. Again, at any point in the experiment the reference machine will have received a random description τ of certain length, and as a result has produced a string of output bits ρ. And, again, what we are interested in, for a specific σ, is the probability that we obtain in this way an output string $\sigma*$ that starts with our σ. This is just the probability

that we arrive at a description for it, a τ such that $U(\tau) = \sigma*$. In this case, since any extension $\tau*$ of a description τ of σ is still a σ-description, we only care about the *minimal* σ-descriptions that are no longer descriptions if we chop one bit off. In short, the probability of obtaining a sequence that starts with σ in this experiment is the sum of the probabilities $2^{-|\tau|}$ for every minimal U-description τ of σ:

Definition 1.2 (Algorithmic Probability [163])**.**

$$Q_U(\sigma) := \sum_{\tau \in T_\sigma} 2^{-|\tau|},$$

with $T_\sigma = \{\tau \mid U(\tau) \succcurlyeq \sigma\}'$ the set of minimal descriptions of σ.

Presupposing the prior selection of a *reference* universal monotone machine, we can drop the subscript U. This defines what would come to be called the *algorithmic probability* $Q(\sigma)$ of given σ: the probability that one obtains it from the reference universal machine on completely random noise.

Description lengths and encodings In the bit-feeding experiment, shorter descriptions naturally have a higher probability, and so strings with shorter descriptions end up with a higher algorithmic probability. This feature is customarily interpreted as a bias towards *simplicity* (Chapter 2, Subsection 2.2.3).

Recall from Subsubsection 1.2.1.2 that the circumstance that descriptions of equal length are *all assigned equal probability* might also be seen to be covered by the principle of indifference. The partition in descriptions *of equal length*, however, would then be a proper addendum to the original principle. The idea that shorter descriptions should receive higher probability does obtain support from another angle: optimal coding in information theory[7].

I already quoted Solomonoff in Subsubsection 1.1.1.2 expressing his debt to the idea that quantity of information is closely related to probability. He specifically mentions the method of Huffman coding (page 12) for lossless data compression. Given a number of symbols with known probability (like the letters of the alphabet, with their relative frequency in a corpus of English texts), this method of coding assigns prefix-free bit strings to symbols in such a way as to attain maximal efficiency or minimal expected codelength. This is achieved, basically, by assigning shorter codes to more frequent or probable symbols. Algorithmic probability turns this idea the other way around: shorter codes are turned into higher probabilities. This intuitive relationship with optimal coding was welcomed by Solomonoff as "strong evidence that I was on the right track" [169, p. 75-6].

[7] It seems proper to note that information theory has also inspired an extension of the original principle of indifference: the *principle of maximum entropy* by E.T. Jaynes [88, 87, 89]. The principle tells us to select the probability distribution for which, within given constraints, the Shannon entropy is highest. Not wanting to digress, I will only say here that there are some strong formal ties with MDL, and refer for further details to [143, 49, 65, 66].

1.2.3.3 Invariance

There remains an obvious worry about the definition of algorithmic probability, indeed about all the foregoing definitions. Just like there exist many different general programming languages, and many, many more possible ones to design, there are an infinite number of ways of defining your favourite universal machine. And here I just stipulated one particular reference machine, with nothing of an argument; a machine that, nevertheless, the definition critically relies on: doesn't this fact make the definition intrinsically subjective?

Recall the Invariance Theorem that I mentioned in the discussion about Kolmogorov complexity on page 24. The result that for any M there is an independent constant c_M such that for all strings σ it holds that $K(\sigma) \leq K_M(\sigma) + c_M$ also implies that for definitions K_1 and K_2 based on universal machines U_1 and U_2 there are c_1 and c_2 such that both $K_1(\sigma) \leq K_2(\sigma) + c_1$ and $K_2(\sigma) \leq K_1(\sigma) + c_2$ for all possible σ. (I will simply write $K_1 =^+ K_2$.) This stems from the fact that a universal machine can emulate any other machine, including other universal machines. The constant represents the additional code length that is needed to describe the machine to emulate. While the additive constants could be big, they are *constant*, which means that as sequences get increasingly long, and their minimal descriptions increasingly large, the additive constants get increasingly insignificant.

Likewise, there holds a straightforward Invariance Theorem for algorithmic probability.

Theorem 1.3 (Invariance Theorem [163])**.** *For two alternative definitions Q_1 and Q_2 of the algorithmic probability (1.2) based on any two universal machines U_1 and U_2, it holds that*

$$Q_1 =^\times Q_2.$$

That is, for these definitions Q_1 and Q_2 there are independent constants c_1 and c_2 such that for any string σ it holds that $Q_1(\sigma) \leq c_1 Q_2(\sigma)$, and also $Q_2(\sigma) \leq c_2 Q_1(\sigma)$.

Again we have that as sequences get increasingly long, and their probabilities increasingly small, the multiplicative constants get increasingly insignificant. Thus the invariance theorem states that asymptotically, the choice of universal machine doesn't make a difference. It's inessential to the theory, in a way that has been likened to the choice of a particular coordinate system in geometry (e.g., [4, p. 3489]). The theorem gives formal support to the objectivity of algorithmic probability.

1.3 The Universal Prior Distribution

The last definition that Solomonoff considered in his 1964 article is quite different from the previous ones. In a sense, the definitions we've looked at in the foregoing Section all took a *bottom-up* approach: they started off from descriptions given to the reference machine, and formalized how the processing of these descriptions gives rise to a distribution on binary sequences. The definition that is the topic of this

Section takes more of a *top-down* approach, by taking into consideration *all possible probabilistic models of induction* from the start.

1.3.1 A Universal Mixture Distribution

One can define a probability distribution by bluntly taking a weighted mean of all possible probability distributions: a *mixture distribution*. This substitutes the original problem for the stipulation of a certain weighting of these distributions, a Bayesian assignment of prior probabilities to the distributions themselves.

However, this still prompts one to decide on the class of possible probability distributions. Again, the key to a successful formalization is to only consider distributions that are *effective* in some specific sense.

1.3.1.1 The Fourth Definition: a Mixture Distribution

Solomonoff bases the prior weighing of distributions on the relative number of descriptions of these distributions. As a description of a distribution P he takes any τ such that $U(\tau\sigma)$ for all given σ produces the binary expansion of the probability $P(\sigma)$. Let $f_{i,n}$ be defined as the fraction of descriptions of the i-th distribution P_i of length n. Then the definition is given by

$$P_{\text{IV}} := \lim_{n\to\infty} \sum_i f_{i,n} P_i(\sigma). \tag{1.8}$$

This definition is not without problems, the most obvious being that the class of *all* probability distributions is way too large to be enumerable, let alone in an effective way. This makes the definition strictly meaningless.

1.3.1.2 The Class of Semicomputable Semimeasures

It was L.A. Levin who put the general idea on a more secure footing, after learning of Solomonoff's paper from Kolmogorov. Levin made an in-depth investigation of the mathematical properties of the class $\mathcal{M}$ of *lower semicomputable semimeasures* [191, 116].

Measures and semimeasures For any given finite string σ, one can define its *cylinder set* as the set $\Gamma_\sigma = \{S \mid \sigma \prec S\}$ of all its infinite extensions. Then I will call a *(probability) measure* a function $\mu : \mathcal{G} \to \mathbb{R}$ from the collection of cylinder sets $\mathcal{G} = \{\Gamma_\sigma \mid \sigma \in 2^{<\omega}\}$ to the real numbers that satisfies

$$\begin{aligned} \mu(\Gamma_\epsilon) &= 1, \\ \mu(\Gamma_\sigma) &= \mu(\Gamma_{\sigma 0}) + \mu(\Gamma_{\sigma 1}). \end{aligned}$$

These conditions can naturally be interpreted as expressing that the probability that an infinite string S is an extension of the empty string (i.e., is in the class of *all* infinite strings) equals 1, and that the probability that S is an extension of σ is the probability that it is

either an extension of $\sigma 0$ or $\sigma 1$. If one identifies the cylinders with their defining strings, one simply gets

$$\begin{aligned} \mu(\epsilon) &= 1, \\ \mu(\sigma) &= \mu(\sigma 0) + \mu(\sigma 1). \end{aligned}$$

For the purpose of this thesis, then, a measure μ is really just a probability distribution P over finite strings that satisfies the natural restrictions that $P(\epsilon) = 1$ and the *subadditivity property* that $P(\sigma) = P(\sigma 0) + P(\sigma 1)$.

A *semimeasure* $\mu : \mathcal{G} \to \mathbb{R}$ only satisfies

$$\begin{aligned} \mu(\Gamma_\epsilon) &\leq 1, \\ \mu(\Gamma_\sigma) &\leq \mu(\Gamma_{\sigma 0}) + \mu(\Gamma_{\sigma 1}). \end{aligned}$$

One can view a strict semimeasure (where some of the above inequalities are strict) as a "defective" probability distribution [50].

Computability of real-valued functions The functions that a Turing machine computes, that I presented as functions on finite bit strings, are really (partial) functions on any countable domain. After all, we can easily define an encoding of a countably infinite set of discrete elements onto finite bit strings. In this way we can treat the computable functions as functions from finite bit strings to natural numbers, or even from rational numbers to rational numbers. What about functions that return real numbers? Real numbers really correspond to *infinite* binary sequences – simply put them in binary (rather than decimal) base.

Much like one can construct the real numbers as Cauchy sequences of rational numbers, one can *computably approximate* a real-valued function by rational-valued (i.e., "finite string-valued") functions. In particular, function $f : 2^{<\omega} \to \mathbb{R}$ is *computably approximated from below* by computable $\phi : 2^{<\omega} \times \mathbb{N} \to \mathbb{Q}$ if $\phi(\sigma, n) \geq \phi(\sigma, n+1)$ and $\lim_{n\to\infty} \phi(\sigma, n) = f(\sigma)$. For any given value function ϕ approaches f ever closer, but one can never tell how far it's still removed. If f can be approximated in this way, it is *lower semicomputable*. Similarly, a function f may be approximated from above by a computable ϕ, in which case it is *upper semicomputable*. A *semicomputable* function is lower *or* upper semicomputable. A function that is *both* upper and lower semicomputable can be nailed down to any degree of precision: it is *computable*.[8]

The foregoing definitions of (semi)computability transfer directly to measures and semimeasures, that are also functions from finite strings to reals.

1.3.1.3 A Universal Element

Again notions of computability provide a natural and very unrestrictive way of coming to grips with concepts of a generality that would simply be too staggering if not tamed at all – in this case a class of all feasible probability distributions. The reason for the weakening to *semi*measures that are only *semi*computable is a feature that is lacked by the subclass $\mathcal{M}^{\text{meas}}$ of computable measures (that coincides with the class of semi-computable measures, because any semicomputable measure must be computable [119,

[8] Here I follow the presentation of [119]. In different contexts, (lower) semicomputable functions are often called (*left-*) *computably enumerable* (*c.e.*) functions (also see [46, ch. 5]). Sometimes they are simply called *enumerable* (e.g., [81]).

Lemma 4.5.1]) and by the subclass of $\mathcal{M}^{\text{comp}}$ of computable semimeasures. Namely, the class $\mathcal{M}$ of semicomputable semimeasures contains formally *universal* members, elements in the same class that *dominate* all others.

Definition 1.4 (Universality of semicomputable semimeasures [191])**.** Semicomputable semimeasure ν is formally *universal* if for every semimeasure $\mu \in \mathcal{M}$ there is a single independent constant c_μ with

$$\nu(\sigma) \geq c_\mu \mu(\sigma)$$

for all bit strings σ.

(I refer to this defining *dominance* property as *formal* universality, to distinguish it from informal conceptions of universality in later discussions.)

A particularly straightforward way of constructing such a universal element exploits the fact that we can effectively enumerate all semicomputable semimeasures $\mu_0, \mu_1, \mu_2, ...$ This allows us to define, in a way that is itself lower semicomputable, a weighted sum of the probabilities given by all measures. For the lower semicomputable *weight function* w, that gives a prior weight to each element of $\mathcal{M}$, it's only required that $w(\mu_i) > 0$ for all i and that the sum $\sum_i w(\mu_i)$ doesn't exceed 1. This leads to a:

Definition 1.5 (Universal Mixture Distribution [191])**.**

$$\xi_w(\sigma) := \sum_i w(\mu_i)\mu_i(\sigma).$$

Stipulating a "reference" function w again permits us to speak about *the* universal mixture distribution. This universal mixture is an example of a (defective) distribution that is formally universal: for every $\mu \in \mathcal{M}$ there is a constant $c_\mu = w(\mu)$ such that $\xi(\sigma) \geq c_\mu \mu(\sigma)$ for all σ. On a more intuitive level, the bounds of formal universality force our ξ to levelly spread out its probability mass over all possibilities, keeping a close distance to each of them.

Incomputability of universal elements Let's refer to the class of all elements that satisfy Definition 1.4 as the class $\mathcal{U}$ of universal elements of $\mathcal{M}$. The definition of universal mixture specifies a specific subclass $\mathcal{U}_\xi \subseteq \mathcal{U} \subseteq \mathcal{M}$ of universal elements of the class of semicomputable semimeasures.

By definition, all the elements of $\mathcal{M}$, including the universal elements, are semicomputable. But the universal elements of $\mathcal{M}$ are certainly not *computable*. This follows most directly from the fact, mentioned above, that the subclass of *computable* continuous semimeasures has *no* universal element [119, Lemma 4.5.2]. Now any universal $\nu \in \mathcal{U}$ certainly dominates all elements in the subclass of computable continous semimeasures: if it were computable, it would be an element, and so a *universal* element, of this subclass. But there exists no such element, hence ν can't be computable.

Figure 1.1 depicts some of the previous facts about $\mathcal{M}$ in a Venn diagram.

1.3.2 Algorithmic Probability and The Universal Mixture

Interestingly, the definitions of algorithmic probability and the universal mixture distribution turn out to be very much related.

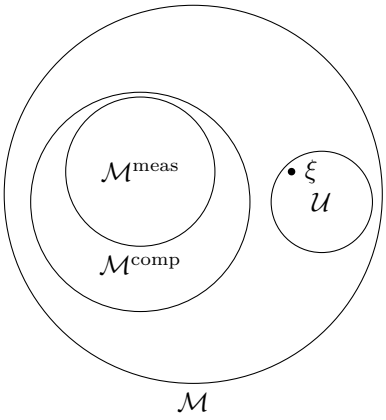

Figure 1.1: The class $\mathcal{M}$ and the relation between subclasses $\mathcal{M}^{\text{meas}}$, $\mathcal{M}^{\text{comp}}$ and $\mathcal{U}$.

1.3.2.1 Algorithmic Probability as a Member of $\mathcal{M}$

Recall that, due to Kraft's inequality, the definition of algorithmic probability based on a prefix universal machine yields a total sum $\sum_\sigma Q(\sigma)$ of probabilities that's bounded by 1. Moreover, closer inspection of the definition shows that the sum of the probabilities $Q(\sigma 0)$ and $Q(\sigma 1)$ of the one-bit extensions of σ is bounded by the probability $Q(\sigma)$ of σ. Hence Q conforms to the conditions of a semimeasure.

Furthermore, one can effectively approximate Q from below. Given a string σ, simply run the reference machine in dovetailing fashion[9] on all possible inputs and put τ in auxiliary set D_σ if the machine outputs σ on reading the input up to τ, and τ is minimal so far (in case there are extensions present already, take them out). At any time, the sum $\sum_{\tau \in D_\sigma} 2^{-|\tau|}$ for provisional D_σ will be bounded by $Q(\sigma)$, never decrease (even if extensions are replaced by a shorter minimal description), and reach $Q(\sigma)$ in the limit. Hence Q is lower semicomputable.

Not a measure, not computable Is Q in fact a proper measure? It's not, and the reason for that is the circumstance that the reference machine doesn't converge on all inputs. For some infinite sequences on the input tape it will never halt. Since one could create a new and larger prefix-free set of descriptions by adding to the reference machine's domain initial segments of these "invalid" sequences (descriptions for "no output", as it were), and still remain within the bounds of Kraft's inequality, the sum of the original probabilities must be strictly below 1. Hence Q is a strict semimeasure.

Moreover, Q is certainly not computable *simpliciter*. If it were, then it would also be possible to compute the length of the *shortest* U-descriptions for each element. To

[9] Let the machine first execute one instruction on the first input, then one instruction on the second input, then two instructions on the first input and two on the second and two on the third, and then three on the first... and so on. This method of *dovetailing* guarantees that every convergent computation will be completed even if the machine encounters divergent inputs.

see this, recall that we can computably enumerate all descriptions of given σ. Now at some point we are sure to know that there can exist no shorter description for σ than the shortest one we have found so far, because an even shorter description would increase the algorithmic probability of σ beyond the value that, we assume, we already have been able to compute. The length of the shortest description we have found in this way is actually the monotone Kolmogorov complexity $Km(\sigma)$ of string σ, the version of Kolmogorov complexity based on monotone machines. However, monotone Kolmogorov complexity, like prefix Kolmogorov complexity, is incomputable – and so, by contradiction, algorithmic probability must be incomputable as well.

Monotone machines and semicomputable semimeasures As a matter of fact, it can be shown that the class of semicomputable semimeasures $\mathcal{M}$ is precisely given by the class of semimeasures defined with the help of a monotone machine (that doesn't have to be universal) as in the definition of algorithmic probability (1.2). That is, for *every* monotone machine M the distribution

$$\mu_M(\sigma) = \sum_{\tau \in D_\sigma} 2^{-|\tau|},$$

with $D_\sigma = \{\tau \mid M(\tau) \succcurlyeq \sigma\}'$ the corresponding set of minimal descriptions of σ, is in fact a semicomputable semimeasure, a member of $\mathcal{M}$. What's more, *all* members of $\mathcal{M}$ are obtainable in this way [191].

1.3.2.2 Asymptotic Equivalence of the Two Definitions

The fact that all universal elements are asymptotically equivalent, together with the fact that algorithmic probability defines a universal element, leads to the asymptotic equivalence with the definition of the universal mixture distribution.

Equivalence of universal elements The choice of weight function w in Definition 1.5 is of no consequence for the dominance properties of ξ. It follows that two universal elements of $\mathcal{M}$, specified in the same way with the help of different semicomputable weight functions, dominate *each other*. This implies, similar to the case of Q and its nonessential reliance on the choice of a reference machine, a specific invariance to the choice of weight function.

For if, suppose, ξ_w were defined with weight function w, and ξ_v with function v, then (and I take $\xi_v = \mu_j \in \mathcal{M}$ and constant $c_v^w = w(\xi_v)$)

$$\xi_w(\sigma) = \sum_i w(\mu_i)\mu_i(\sigma) \geq w(\mu_j)\mu_j(\sigma) = w(\xi_v)\xi_v(\sigma) = c_v^w \xi_v(\sigma)$$

for all strings σ. Likewise, there is a constant c_w^v such that $\xi_v(\sigma) \geq c_w^v \xi_w$ for all σ. In short, $\xi_w =^\times \xi_v$.

In effect, Definition 1.5 illustrates just one way of constructing a semicomputable semimeasure that fits the requirement of Definition 1.4. Clearly, the relevant trait

for asymptotic equivalence is the abstract property of formal universality, the property of dominance. In general, for every pair ν and ν' of universal semicomputable semimeasures,

$$\nu =^{\times} \nu'.$$

Thus all elements of $\mathcal{U}$ are asymptotically equivalent.

Algorithmic Probability as Universal Element It's not hard to verify that the semimeasure $Q = \mu_U$, based on a reference universal machine that emulates any other machine, dominates all other μ_M. That means that Q is a *universal* element in the class $\mathcal{M}$ of semicomputable semimeasures.

To be a bit more precise, every definition of algorithmic probability Q_U via a specific reference universal machine U specifies a universal element of $\mathcal{M}$. I will denote this subclass of the class $\mathcal{U}$ of universal elements by $\mathcal{U}_Q$.

The universality of Q entails the (on the face of it) remarkable asymptotic equality of the algorithmic probability Q and the universal mixture distribution ξ.

Theorem 1.6 (Asymptotic equivalence Q and ξ [191])**.** *There exists a constant c such that for each $\sigma \in 2^{<\omega}$:*

$$Q(\sigma) = \xi(\sigma)$$

up to multiplication by c.

That is, $Q(\sigma) =^{\times} \xi(\sigma)$. Equivalently, $-\log Q(\sigma) =^{+} -\log \xi(\sigma)$. In more precise terms this means that every element of $\mathcal{U}_Q$ is asymptotically equivalent to every element of $\mathcal{U}_\xi$:

$$Q_U(\sigma) =^{\times} \xi_w(\sigma)$$

for every universal machine U and semicomputable weight function w.

1.3.2.3 The Universal Prior Distribution

The equivalence of both definitions gives us ground to take the convenient step of pretending that we have one unique universal distribution.

The class of universal semicomputable semimeasures The two definitions of algorithmic probability and the universal mixture distribution are equivalent in yet another strong sense, as was very recently shown by I. Wood, P. Sunehag and M. Hutter [188]. Namely, any universal semimeasure defined as the algorithmic probability via some universal U is *equal* (not just asymptotically equivalent) to the universal mixture distribution via some semicomputable weight function, and vice versa. That is, both definitions give rise to the exact same class of universal semicomputable semimeasures.

Theorem 1.7 (Relation definitions Q and ξ [188])**.** *The class $\mathcal{U}_Q$ of all algorithmic probability distributions Q_U coincides with the class $\mathcal{U}_\xi$ of all universal mixture distributions ξ_w:*

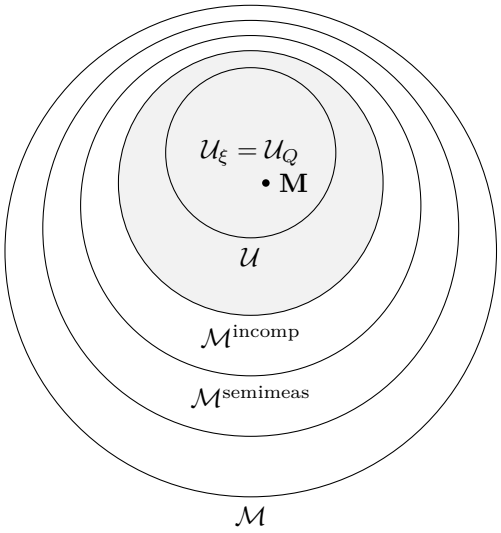

Figure 1.2: The constitution of the class $\mathcal{M}$ of semicomputable semimeasures. $\mathcal{M}^{\text{semimeas}} = \mathcal{M} \setminus \mathcal{M}^{\text{meas}}$ represents the subclass of strict semimeasures and $\mathcal{M}^{\text{incomp}} = \mathcal{M} \setminus \mathcal{M}^{\text{comp}}$ the subclass of incomputable elements. All elements in the shaded area are asymptotically equivalent.

$$\mathcal{U}_Q = \mathcal{U}_\xi.$$

In the same paper, the authors show that not all universal semicomputable semimeasures are of this form: the class $\mathcal{U}$ strictly includes the class $\mathcal{U}_Q = \mathcal{U}_\xi$.

Theorem 1.8 (Relation $\mathcal{U}$ and $\mathcal{U}_Q = \mathcal{U}_\xi$ [188])**.** *There exist universal semicomputable semimeasures that are not given by the universal probability via some universal machine U (equivalently, the universal mixture via some semicomputable weight function w):*

$$\mathcal{U}_Q = \mathcal{U}_\xi \subsetneq \mathcal{U}.$$

The picture that emerges from these results is given in Figure 1.2.

The universal prior distribution There is as yet little known about the properties of universal elements outside the class $\mathcal{U}_Q = \mathcal{U}_\xi$, and in what respects they might differ from elements that can be defined as universal mixture or algorithmic probability distributions [M. Hutter, personal correspondence]. In all of the following, I will restrict attention to those universal elements that are covered by the two equivalent definitions.

Within this class, the equivalence of Theorem 1.7 tells us that the selection of a universal mixture ξ_w by the choice of a reference weight function w comes down to the selection of a reference universal monotone machine U to define an algorithmic

probability distribution Q_U, and the other way around. In the upcoming discussions, it will be convenient to suppose that such a selection has been made, resulting in a unique distribution $\xi_w = Q_U$ that is denoted $\mathbf{M}$. Alternatively, in light of the asymptotic equivalence of all of these distributions, it sometimes makes practical sense to view all elements of $\mathcal{U}_Q = \mathcal{U}_\xi$ as essentially representing one single distribution, again denoted $\mathbf{M}$. In both ways of looking at things, I will refer to this distribution $\mathbf{M}$ as the *universal prior distribution*.

At other times, it's important to stress the fact that in reality there is an infinite class $\mathcal{U}_Q = \mathcal{U}_\xi$ of asymptotically equivalent predictors. Overloading the previous terminology, I baptise this class $\mathcal{U}_\mathbf{M}$, the class of universal prior distributions.

1.4 Universal Prediction

The problem that we started with was to predict the continuation of a binary string. I conclude this Chapter with the solution according to Solomonoff's theory of Prediction.

1.4.1 Prediction with Solomonoff

One point of departure was that the developing history is governed by a probability distribution. The other is that the development is effective. More precisely, putting things in Levin's framework, the developing string is governed by a computable measure. Since an effective measures is supposed to give a complete specification of the bounds on the data generation process, I will often figuratively use the term *environment* to refer to any such measure.

We suppose there is always an underlying "true" or actual measure $\mu \in \mathcal{M}^{\text{meas}}$ at work in generating the string in question. Falling back on Bayes' rule (1.3), the actual probability of τ after having seen σ would then be

$$\mu(\tau \mid \sigma) = \frac{\mu(\sigma\tau)}{\mu(\sigma)}. \tag{1.9}$$

Of course, the problem is that we can't assume to know the actual environment μ. Therefore, instead of the actual μ, we use the universal prior distribution $\mathbf{M}$. Replacing the unknown μ with $\mathbf{M}$, we estimate the probability of continuation τ of σ as

$$\mathbf{M}(\tau \mid \sigma) = \frac{\mathbf{M}(\sigma\tau)}{\mathbf{M}(\sigma)}. \tag{1.10}$$

Predictors A natural way of looking at the universal prior distribution is as a universal method of prediction – the *universal predictor* $\mathbf{M}$. In most part of the thesis, I refer to $\mathbf{M}$ in this way. Again, where this can't lead to trouble, I will often follow custom in pretending there is a single such predictor, rather than an infinite class $\mathcal{U}_\mathbf{M}$ of asymptotically equivalent ones.

The unknown environment μ could also be seen as representing the most "informed" prediction method available – we could call it the *informed predictor*. However, with

the important exception of Chapter 4, Section 4.4, I will stick to the customary interpretation as an actual generation mechanism of the data, an actual environment.

Measures The class $\mathcal{M}$ that constitutes the formal framework consists of all semicomputable *semi*measures, and the universal predictor in the form of the universal mixture distribution takes every such element into account. Yet we limit ourselves to the *measures* (for which semicomputability and computability coincide) as the class of possible actual distributions. This issue will receive further attention in Chapter 4, Section 4.2, but I will say here that one important reason for the limitation to measures is the central result of the next Subsection.

1.4.2 Completeness

So what is the rationale for this strategy, for always using the universal predictor $\mathbf{M}$? There is, to begin with, the intuitive support of the ideas behind the definition of Q and ξ, and of the domination properties of the universal prior distribution as universal element. I will come to deal with the weight of these intuitions and underlying assumptions in detail in the upcoming Chapters. But there is more to offer than just the hopeful fancy that applying the universal prior instead of the actual distribution for prediction will lead to good results: it can be rigorously *proven*. With good results I mean that the predictions with $\mathbf{M}$ will very quickly be close to the probabilities given by the actual distribution. Our predictions will rapidly *converge* to the best possible predictions.

1.4.2.1 Convergence in Difference

A suitable measure of the expected error of the n-th prediction when using $\mathbf{M}$ instead of the actual μ is given by

$$H_t(\mu, \mathbf{M}) := \sum_{\{\sigma:|\sigma|=t-1\}} \mu(\sigma) H(\mu(\cdot \mid \sigma), \mathbf{M}(\cdot \mid \sigma)),$$

with

$$H(\mu(\cdot \mid \sigma), \mathbf{M}(\cdot \mid \sigma)) := \sum_{b\in\{0,1\}} (\sqrt{\mu(b \mid \sigma)} - \sqrt{\mathbf{M}(b \mid \sigma)})^2$$

the squared *Hellinger distance* between μ and $\mathbf{M}$ (conditional on σ). The following Theorem states the convergence of the total sum, for all lengths t, of all expected errors after t bits.

Theorem 1.9 (Convergence in difference [165])**.** *For (computable) measure μ, the expected total prediction error when using $\mathbf{M} = \xi_w$ is*

$$\sum_t H_t(\mu, \mathbf{M}) \leq -\ln w(\mu).$$

Proof. The proof I give here is based on [119, Theorem 5.2.1]. I will actually show this bound for a different measure of expected error, namely

$$D_t(\mu \parallel \mathbf{M}) := \sum_{\{\sigma:|\sigma|=t-1\}} \mu(\sigma) D(\mu(\cdot \mid \sigma) \parallel \mathbf{M}(\cdot \mid \sigma)),$$

with

$$D(\mu(\cdot \mid \sigma) \parallel \mathbf{M}(\cdot \mid \sigma)) := \sum_{b \in \{0,1\}} \mu(b \mid \sigma) \ln \frac{\mu(b \mid \sigma)}{\mathbf{M}(b \mid \sigma)}$$

the *Kullback-Leibler divergence* of $\mathbf{M}$ with respect to μ (conditional on σ).[10] The original theorem then follows immediately because $H_t \leq D_t$ for all t (see [119, Lemma 5.2.1, Claim 5.2.1]).

We can derive the upper bound $-\ln w(\mu)$ on the sum $\sum_{t=1}^{s} D_t$ as follows.

[10] The Kullback-Leibler divergence, also known as *relative entropy*, is an information-theoretic way of quantifying the error of one distribution from the perspective of another. Unlike the Hellinger distance, the Kullback-Leibler divergence is not a metric, which is sometimes taken as cause to prefer the Hellinger distance as a measure of difference between distributions.

$$
\begin{aligned}
\sum_{t=1}^{s} D_t(\mu \parallel \mathbf{M}) &\overset{(a)}{=} \sum_{t=1}^{n} \sum_{|\sigma|=t-1} \mu(\sigma) \sum_{b} \mu(b \mid \sigma) \ln \frac{\mu(b \mid \sigma)}{\mathbf{M}(b \mid \sigma)} \\
&\overset{(b)}{=} \sum_{t=1}^{s} \sum_{|\sigma|=t-1} \sum_{b} \mu(\sigma b) \ln \frac{\mu(b \mid \sigma)}{\mathbf{M}(b \mid \sigma)} \\
&\overset{(c)}{=} \sum_{t=1}^{s} \sum_{|\sigma|=t} \mu(\sigma) \ln \frac{\mu(\sigma_t \mid \sigma_{<t})}{\mathbf{M}(\sigma_t \mid \sigma_{<t})} \\
&\overset{(d)}{=} \sum_{t=1}^{s} \sum_{|\sigma|=s} \mu(\sigma) \ln \frac{\mu(\sigma_t \mid \sigma_{<t})}{\mathbf{M}(\sigma_t \mid \sigma_{<t})} \\
&\overset{(e)}{=} \sum_{|\sigma|=s} \mu(\sigma) \ln \prod_{t=1}^{s} \frac{\mu(\sigma_t \mid \sigma_{<t})}{\mathbf{M}(\sigma_t \mid \sigma_{<t})} \\
&\overset{(f)}{=} \sum_{|\sigma|=s} \mu(\sigma) \ln \frac{\mu(\sigma)}{\mathbf{M}(\sigma)} \\
&\overset{(g)}{\leq} \sum_{|\sigma|=s} \mu(\sigma) \ln \frac{\mu(\sigma)}{w(\mu)\mu(\sigma)} \\
&\overset{(h)}{=} \sum_{|\sigma|=s} \mu(\sigma) \ln w(\mu)^{-1} \\
&\overset{(i)}{\leq} -\ln w(\mu).
\end{aligned}
$$

Step (a) is just the unfolding of the definition of D_t. Step (b) brings $\mu(\sigma)$ into the sum and then uses $\mu(\sigma b) = \mu(\sigma)\mu(b \mid \sigma)$.

In step (c), we merge the summation over single extensions b by summing over strings σ that are one bit longer. Here I use the shorthand σ_t to denote $\sigma(t)$, the t-th bit of σ, and $\sigma_{<t}$ to denote the initial segment $\sigma \restriction_t$ of σ.

Step (d) exploits the equality $\mu(\sigma_{<t}) = \sum_{|\sigma|=s} \mu(\sigma)$ that is the strict subadditivity of measures. This step thus requires μ to be a proper measure. The result of (d) is that we now always sum over the strings σ of length s, for each t of the outermost sum. That means we may move the $\sum_{t=1}^{s}$ to the right, leading to a sum of logarithms that may be rewritten as the logarithm of a product: this is step (e). By the trick of *telescoping* this product of conditionals is reduced in step (f) to the simple fraction $\mu(\sigma)/\mathbf{M}(\sigma)$.

At this point, we apply the essential formal universality of the universal mixture:

$$\mathbf{M}(\sigma) = \xi_w(\sigma) = \sum_i w(\mu_i)\mu_i(\sigma) \geq w(\mu)\mu(\sigma)$$

to derive to inequality of step (g). The resulting term is then further simplified via (h) and (using that the sum of probabilities of strings of the same length is bounded by 1) via (i), finally resulting in the bound $-\ln w(\mu)$. □

Thus the total expected sum of prediction errors in the limit is bounded by a small constant $k = -\ln w(\mu)$, that only depends on the weight of μ (via the weight function w used in defining ξ_w). The convergence of the total sum implies that, in the limit of t, the expected error H_t after t bits goes to 0. Even better, the *tail sum* $\sum_{t=s}^{\infty} H_t$ also converges to 0 as s goes to the limit.

The Theorem entails that the universal predictor $\mathbf{M}$ can be expected (with probability 1) to quickly converge to any unknown distribution.

Corollary 1.10. *For any (computable) measure μ,*

$$\mathbf{M}(b \mid \sigma \upharpoonright_t) - \mu(b \mid \sigma \upharpoonright_t) \overset{t\to\infty}{\longrightarrow} 0$$

with μ-probability 1.

Convergence in Ratio The previous result doesn't exclude the possibility that the *ratio* between the values $\mathbf{M}(b \mid \sigma)$ and $\mu(b \mid \sigma)$ keeps on fluctuating and never converges. The following theorem addresses this worry.

Theorem 1.11 (Convergence in ratio [P. Gács], see also [119, 81])**.** *For (computable) measure μ and infinite sequence $S \in 2^\omega$,*

$$\frac{\mathbf{M}(S(t) \mid S \upharpoonright_t)}{\mu(S(t) \mid S \upharpoonright_t)} \overset{t\to\infty}{\longrightarrow} 1$$

with μ-probability 1.

This result only concerns *on-sequence* convergence, keeping to the probabilities of bits $S(n)$ on the sequence S. For *off-sequence* prediction in general, the additional requirement is needed that all conditional μ-probabilities stay away at least some fixed constant from 0.

Further convergence and optimality results Over the last decade, Hutter has further refined these convergence results. He generalized them to arbitrary finite alphabets [76] and added some facts about the speed of convergence [78, 86].

In addition, he inferred a tight upper bound on the loss suffered by the universal predictor in comparison to the (optimal) informed predictor for a wide variety of loss functions [77, 78]. Moreover, he showed that the universal predictor is *Pareto-optimal*, meaning that no predictor can do better than the universal one in every environment [79].

1.4.2.2 Completeness

The result that the universal predictor quickly converges to any unknown distribution, is what Solomonoff himself brands the *completeness* of his method. He presents it as standing in an inverse relation to computability: if "probability evaluation methods"

are "complete, and hence 'uncomputable', they have unknown error size. If they are computable, then they *must* be incomplete." [167, p. 474]

This Section's results provide support for the power of $\mathbf{M}$ as a method of prediction, converging quickly to any measure in $\mathcal{M}$. At the same time, $\mathbf{M}$ is itself a semimeasure in $\mathcal{M}$: the universal prior distribution. As the name suggests, it could equivalently be viewed as a candidate distribution for assigning the priors, supported by the same results. Thus Li and Vitányi conclude:

> "The problem with Bayes' rule has always been the determination of the prior. Using $\mathbf{M}$ universally gets rid of this problem and is provably perfect." [119, p. 361]

2 The Principles of Solomonoff Prediction

In the presentation of Li and Vitányi, the theory of Solomonoff arises from a small number of core ideas or principles:

> "Essentially, combining the ideas of Epicurus, Ockham, Bayes and modern computability theory, Solomonoff has successfully invented a perfect theory of induction. It incorporates Epicurus's multiple explanations idea, since no hypothesis that is still consistent with the data will be eliminated. It incorporates Occam's simplest explanation idea, since the hypotheses with low Kolmogorov complexity are more probable. The inductive reasoning is performed by means of the mathematically sound rule of Bayes." [119, p. 347]

The overview of the previous Chapter was mainly aimed at the formal aspects of Solomonoff's theory of Prediction. In this Chapter, I shift focus to the proper way of looking at the formal theory, of *interpreting* the theory. This, firstly, requires a consideration of what we actually take to be the purpose or aim of the theory, which will gain us a clearer view on what we should require of it (Section 2.1). This discussion, secondly, motivates and supports the identification of core principles that can be said to underly the theory (Section 2.2).

Although it is initiated by the proposed main ideas in the above quotation, the discussion of the role and status of the candidate core principles will lead me to take a critical view on the way SP is often presented in the literature. This clears the way for an alternative interpretation in terms of the one core principle of *Universality*.

2.1 The Purpose of Solomonoff Prediction

In order to get started, we have to agree on what we actually require of Solomonoff Prediction. What is it supposed to give us, by what standards should we judge it? It is the question of the purpose of Solomonoff's theory and its resulting goals that initiates our quest in Section 2.2 for the core principles of SP.

In this Section, I will first distinguish three ways of interpreting a project like Solomonoff's (Subsection 2.1.1): as a *method*, as a *model*, and as a *theory*. The method interpretation, centered on the universal predictor, is perhaps the dominant view (Subsection 2.1.2). The model interpretation resolves around the issue of the connection of the framework with actual predictive problems (Subsection 2.1.3). I will reject these

first two interpretations in favour of the theory interpretation, that will finally lead me to point out the relevance of identifying the core principles of SP (Subsection 2.1.4).

2.1.1 Method, Model and Theory

As an illustrative prelude to the discussion of the purpose of Solomonoff Prediction, let us first examine in some detail the aims of the more general Bayesian theory of scientific inference (Chapter 1, Subsubsection 1.1.2.2).

2.1.1.1 Bayesian Confirmation as a Theory of Scientific Inference

Bayesian confirmation theory is adopted in the philosophy of science as providing a theory of scientific reasoning. The question to ask is: what should such a theory give us?

Method On a first glance, the Bayesian aim is to offer us a guideline of how rational inference is to be done. The procedure of Bayesian reasoning is just that: a *procedure*, a *method*. And rather than a *descriptive* portrayal of a working scientist's actual reasoning, it's a method with a *normative* flavour: a method of optimally rational, *ideal* reasoning. It must in its pure form be an idealization and not in any sense a *practical* method, one is forced to admit, since in reality already the computational power required to meet the probability axioms (e.g., recognizing all tautologies) is beyond anyone's reach. This raises the obvious question about the use of a seemingly impossible prescription without guidelines how to actually act upon them.

Model There exists a broader interpretation. Indeed, the attraction of the Bayesian approach stems from its success, as a particularly simple yet precise paradigm based on plausible and very mild assumptions, of embedding and defusing a respectable number of practices and problems in (the philosophy of) science; a success that no competitor can boast of. In this guise, Bayesianism does have a descriptive ring: it provides a *model* of scientific practice and problems. It hands a way of representing large part of the scientific enterprise at a level of abstraction that often makes us see things more clearly. Then the level of idealization is directly related to the level of abstraction that is required to actually accommodate a real-world problem. And the *model* can be seen to incorporate the *method*, again as a sufficiently abstract depiction of reasoning.

Still, the bare fact of a model that can accommodate whatever we want is perhaps not that interesting. In the Bayesian model, many particular inferences could be taken apart and accommodated in the model by an *ad hoc* choice of degrees of belief, leading to the worry that "the Bayesian apparatus is just a kind of tally device used to represent a more fundamental sort of reasoning whose essence does not lie in the assignment of little numbers to propositions in accord with the probability axioms" [47, p. 59]. As C. Glymour puts it in a related criticism [60], Bayesian inference then rather resembles a theory of personal *learning*: but showing how I arrive at certain beliefs from prior probabilities doesn't compel anyone else to the same reasoning –

that is, provides no *argument* (see Chapter 3, Subsubsection 3.1.1.1 for an extension of this critique that is relevant to SP).

Theory The above concerns have to do with the *explanatory* value of the model. Bayesian confirmation theory has to address this issue by going beyond just the model. As a *theory*, it consists of statements *about* the model. These statements concern what the model can tell us, and may also include justifications of the rationality and objectivity of the model, and of the method that the model incorporates. Showing how I arrive at certain beliefs from prior probabilities *could* turn out to provide an argument if the constraints on the prior probabilities can be seen to stem from more fundamental principles, and these ought to be found in the complete theory. In short, the theory specifies the conclusions that can be drawn from the model and method, and, ideally, also suggests *why*.

Alternative interpretations The final aim of the theory of Bayesian confirmation as a whole would be no less than a completely general and unified explanation of scientific method. Unfortunately, the great amount of idealization involved, and also the relative failure of the theory's *a priori* arguments for the objectivity and rationality of the model and method, advocates caution in accepting the theory as a full explanation of scientific inference. As a consequence, much effort has been directed at adjusting the theory towards a trade-off between generality and applicability. In a way, this comes down to finding different ways of looking at the same body of results.

If one opts to aim for more modest goals than a complete and systematic explanation of scientific method, one could put the focus back on restricting the reach of the method and/or the model. W.C. Salmon [148], for instance, suggests that (following Kuhn [109]) actual science is about choosing between a theory and its rivals – thus the issue becomes how well *in comparison* with its rival(s) a hypothesis is confirmed by a piece of evidence. In his *Bayesian algorithm for theory preference*, evidently a proper method that is to model this way of doing science, we only have to calculate the ratio of the posterior probabilities of two rival theories, so that, conveniently, the troublesome expectedness of the evidence cancels out. Naturally, this relative confirmation does mean that we lose all absolute estimates of the degree of belief in hypotheses.

A very different approach, following P. Horwich's *therapeutic* or *Wittgensteinian Bayesianism* [72], is to take a view of Bayesianism as what he calls a *philosophical* theory that is primarily aimed at solving puzzles that arise from reflecting on scientific method. The more fruitful approach, Horwich argues, is a *problem*-oriented rather than a theory-oriented philosophy of science, where we are satisfied with dissolving paradoxes by unravelling misguided thinking; as such, Bayesianism is an ideal compromise between accuracy and simplicity, and stubbornly judging it by the criteria of a complete *scientific* theory would be "misplaced scientism".

The previous examples show that there are multiple different ways of making sense of the same theory. Much the same applies to the theory of Solomonoff, as we will see.

2.1.1.2 Method, Model or Theory

The general moral I would like to draw from the previous discussion of the nature of Bayesian confirmation theory is the loose but useful distinction between a *method*, a *model*, and a *theory*.

- By a *method* I mean a well-defined procedure or strategy that applies in a preconceived setting. Roughly, a method describes or prescribes *how* things can or should be done.
- By a *model* I mean a description or depiction of a setting that is supposed to represent part of reality, at some level of abstraction. One could define methods to apply in this setting. Roughly, a model captures *what* we think is relevant.
- By a *theory* I mean an overarching collection of statements that are supposed to have explanatory value. These statements are likely to concern a model that is adopted by the theory. They could express conclusions drawn from the model, and reasons for its validity. Roughly, a theory ought not only to tell us what is the case, but also *why* this is so.

The method, model and theory of Solomonoff Prediction It will be very helpful to also consider the purpose of SP from the perspective of this schema. In the upcoming Subsections, I will consider interpretations of SP as a method of prediction (Subsection 2.1.2), as a model of prediction (Subsection 2.1.3), and as a theory of prediction (Subsection 2.1.4), respectively.

2.1.2 The Method of Solomonoff Prediction

Looking back at the presentation in Chapter 1, Section 1.4, the essence of Solomonoff's theory seems to be the guideline to use the universal predictor $\mathbf{M}$ in all cases. On this *method interpretation* op SP, the whole endeavor revolves around the specification of a particular *method* of prediction.

2.1.2.1 A Ready-To-Apply Method of Prediction

The first thing to get out of the way, then, is the issue of the *practical* use of this method. Even presupposing the restricted setting of effective binary sequence prediction, can it be directly implemented to do actual prediction? The answer shouldn't come as a surprise: it is not, it cannot.

Incomputability Much has been made of the fact that universal predictor $\mathbf{M}$ is formally *incomputable* (see Chapter 1, Section 1.3.2): there can be no algorithm that calculates the $\mathbf{M}$-probabilities of bit strings. This appears to be a clear-cut mathematical result that immediately dashes any hope of practical application.

I have two comments to this rigid reading. First, semimeasure $\mathbf{M}$ is still lower *semi*-computable. We can, in principle, close in on its values from below to any degree

of accuracy, although the incomputability implies that we can never know how close we are. Second, in practice, of course, it doesn't really matter whether a predictor is completely infeasible to calculate or provably incomputable; in both cases, the only way out is some approximation.

2.1.2.2 An Ideal Method of Prediction

Notwithstanding the fact that the universal predictor can never be applied in practice, it's customary in the literature to assess Solomonoff's theory as an actual method of prediction. It's viewed as an *idealized* method, a supposedly optimal but ethereal procedure that sets a paradigm for practical methods. In Solomonoff's own words, "algorithmic probability can serve as a kind of 'gold standard' for induction systems" [169, p. 83]. Again, an ideal that other methods can try to *approximate*.

Naive approximations The most straightforward approximation simply puts a limit on the computation time, or number of computation steps, that we spend on calculating **M**-probabilities. We use the value that we have when time is up.

To illustrate, consider a *resource-bounded* version of algorithmic probability

$$Q^t(\sigma) := \sum_{\tau \in T^t_\sigma} 2^{-|\tau|},$$

with $T^t_\sigma = \{\tau \mid U^t(\tau) \succcurlyeq \sigma\}'$ the set of minimal descriptions of σ found within the time bound given by t. Here $U^t(\tau)$ means that U will halt after $t(|\sigma|)$ steps, for some function t, at the latest.

This method of checking the infeasibility of **M** was already hinted at in Solomonoff's 1964 paper; some versions (also *memory*-bounded) were first formalized in detail in [185]. Solomonoff has high hopes for its use as a "general solution to the problem of *practical* induction" [168, p. 1].

I find it hard to see how we are much better off, for all practical purposes, with this solution. Apart from the inelegance of subjecting the computation to some arbitrary preconceived upper bound, this method of calculation is still as brute-force and inefficient as ever: we just give up at some point. The longer we can suspend this point, the closer we get to the optimal value given by **M**; but we would still need a completely infeasible amount of resources to get anywhere close – and never know how close.

Beyond method Anticipating more extensive discussion in Chapter 3, Section 3.1 on the principle of Completeness, I think that the description of an idealized method, in isolation, has little use. We need more context to make sense of the virtue of a method that is not implementable and that also can't be straightforwardly approximated. This context is provided by an actual theory that embeds this method. As we will yet see, the dominant inclination to keep to the level of this method interpretation, focussing on problems that would only arise if we try to actually employ the method, gives rise to unnecessary problems (see, as a prime example, Chapter 3, Section 3.2).

2.1.3 The Model of Solomonoff Prediction

The universal prediction method $\mathbf{M}$ is embedded in the precise mathematical framework that was developed in the previous Chapter. This mathematical setting, as a representation of predictive problems and the process of predicting, is the model that the theory rests on. On the *model interpretation* of SP, the virtue of the project is just this provision of an abstract model of prediction.

2.1.3.1 The Mathematical Framework

The mathematical setting is given by the class $\mathcal{M}$ of semicomputable semimeasures. One element from its subclass $\mathcal{M}^{\text{meas}}$ of computable measures is supposed to govern a developing binary sequence. The continuation of this sequence may be predicted by the method that is given by a universal element in $\mathcal{U}_{\mathbf{M}} \subset \mathcal{M}$.

The choice for an information-theoretic (Chapter 1, Subsection 1.1.1, probabilistic (Chapter 1, Subsection 1.1.2) and effective (Chapter 1, Subsection 1.1.3) setting (that is subsequently formalized as involving semi*measures* that are semi*computable*) is the point of departure for the model and SP in general. It forms a starting assumption without which the theory wouldn't get off the ground – though I will show how it derives from a more fundamental consideration in Chapter 4.

Some intuition A vivid way of looking at the mathematical framework is by means of *binary trees*. Each (semi)measure μ corresponds to a binary tree where the root node represents the empty string, and every node branches into two nodes that represent the two possible one-bit extensions $\sigma 0$ and $\sigma 1$ of the string σ represented by the original node. The edge leading from a string to a one-bit extension σb is marked by the probability value $\mu(\sigma b)$ of this extension, according to the semimeasure μ.

Now the generation of a bit sequence that is governed by μ can be seen as the undertaking of a journey that departs from the root of the corresponding tree (let's call it the μ-tree), when all you have is an empty sequence. Then at each split you encounter, the probability values at the signposts next to the two roads determine the probability of you taking the left or the right turn. At each such turn, you add the corresponding bit to your sequence.

The procedure of conditionalizing, assessing the relative μ-probabilities after having received string σ, comes down to taking the subtree with root σ from the μ-tree, and normalizing the probabilities at the edges of this subtree (in such a way that the probabilities at the edges from the root sum to 1).

Naturally, one of these binary trees represents the universal predictor. (Or rather, each universal semimeasure is also represented by some such tree.) It has the remarkable property that as given data sequences get increasingly longer, the probability values of the normalized subtrees rooting in these data sequences increasingly agree with those of the corresponding normalized subtrees of the unknown actual μ-tree.

2.1.3.2 The Framework as a Model

The mathematical framework is, of course, highly artificial. It has all the looks of leading an isolated, Platonic existence. The framework can only be commended as an

actual *model* of prediction if it can be tied to the concrete world, if it can be seen to represent a relevant part of the predictive problems we are interested in.

Making the connection to reality Ultimately, we would hope that the framework can be shown to be able to capture predictive problems in the "real world". We would hope that binary sequences can serve as representations of the real-world phenomena that we investigate, and that effective (semi)measures can serve as representations of how they have developed and will develop in the future, and how this future is predicted.

I will, indeed, argue in Chapter 4 that the framework is sufficiently unrestrictive to allow us to accommodate any real predictive problem, in principle. (Note that there is some unclarity about the dual role of (semi)measures as environments and as predictors. This issue will be addressed in Section 4.4 of Chapter 4.)

Beyond model However, the qualifier "in principle" is significant. The downside of the extreme generality of the framework is its extreme level of abstraction. This puts doubt on the suitability of actually modeling given predictive problems.

It certainly looks utterly infeasible to take the trouble to translate a specific predictive problem in terms of binary sequences and effective semimeasures. Even if one persists and completes the job, at an appropriate level of abstraction, it's unclear how this new description would be very informative. What it gives is the opportunity to bring the universal predictor into the problem, but in light of the infeasibility of the method it's again unclear how that would be useful: not as a *descriptive* component because prediction wouldn't be done in that way, nor as a *normative* component because it could in fact never be done in that way. Clearly, the translation from an actual predictive problem to the mathematical framework simply involves too great an amount of idealization to be helpful with the problem at hand.

This consideration defeats the purpose of SP as just providing a model of prediction; hence takes down the model interpretation. The conclusion that it doesn't make sense to handle specific predictive problems via the model, however, doesn't mean that it's not interesting that any predictive problem could, in principle, be accommodated in the model. The interest then lies in what can be said *about* this extremely general model. This is the content of the encompassing *theory* of Solomonoff Prediction.

2.1.4 The Theory of Solomonoff Prediction

The theory of SP is *about* the idealized model and *about* the idealized universal method. What I will come to contend is that the essence of Solomonoff's theory of Prediction is the nontrivial result that *in a universal framework, there exist universal predictors*. What I mean by this, and why I think this is the essence, will become fully clear in the next Chapter on the principle of Universality.

For now, the important thing is that Solomonoff Prediction should be seen as giving a theory of prediction. The ambitious label "theory" suggests a more sweeping status than "just" a method or "just" a model, but it's perhaps more accurate to see the theory interpretation as cautiously taking a step back. Much less exciting than the

unconditioned launch of a single all-purpose predictor, or an all-encompassing model of prediction, as the method and model interpretation have it, the above statement, the content of the theory, expresses a qualification of the significance of the model and method. As such, the theory interpretation of SP that I defend here is really a much more modest interpretation.

The next step is to further explore what we hope to really gain by such an idealized theory – idealized because of the idealization involved in the model (and method) it incorporates. It's the answer to this question that reveals the relevance of the project of this Chapter: the establishment of the core principles of Solomonoff's theory.

2.1.4.1 The Use of an Idealized Theory

I would argue that, in general, the potential merits of an idealized theory are:

- allowing a more profound *understanding* of the relevant issues;
- providing a basis for *approximation* by, or *inspiration* to more practically applicable theories.

These two points are very much connected.

Insight and understanding The least that can be said of an idealized theory like SP is that it touches on some fundamental philosophical problems. The further hope, that we may codify as a further *aim*, is that it allows us to find a new and clearer way of looking at elements of these problems, and to gain a more profound understanding of them.

This goal is still a difficult thing to defend. Faced with a skeptic who dismisses work on extremely idealized theories of this kind as sterile academic occupational therapy, the burden of proof would lie on us. The obvious line of defence against such a skeptic would be the promise of future more practical use.[1] If we take one step in that direction, we arrive at the following point.

Inspiration and approximation As an idealized theory, SP can still set a direction for more practically oriented research. Methods or principles of prediction could be developed that derive from the idealized, strictly unreachable, theory.

Of course, the main example that I have in mind of such a theory is J.J. Rissanen's principle of Minimum Description Length.

This defence also brings to mind Solomonoff's wording of a "gold standard" for induction systems. Then doesn't it pertain to a method interpretation of SP again, as a specific method that others must try to approximate? It's certainly a good guess that this is what Solomonoff had in mind. However, this way of viewing things can only be said to apply to the most naive approximations, the main class of which

[1] Though I believe there is an equally important yet harder to defend intrinsic value to (the development of) such theories, and so accepting the skeptic's criterion, *de facto* accepting to play by his rules, is rather like an admission of weakness. An argument to that effect, however, would take us too far here, and, anyway, the second point suffices for the current exposition.

I discussed and discarded in Subsection 2.1.2 above. Any other practically feasible theory or method will be sufficiently different from the ideal method that many of its aspects can only be said to be *inspired* by the original, rather than being actual *approximations*. Admittedly, these qualities are two ends on the same scale, on which every more down-to-earth theory that aspires to the idealized setting of SP has to find a place. But granted that most interesting theories (including MDL) will tip the balance to the "inspiration" end, the aspects of the idealized method that such a theory will derive inspiration from are not *that* it gives optimal predictions – it will be in aspects that elucidate *why* it gives optimal predictions. This is the burden of the theory that includes the method, and the core principles that underly the theory.

2.1.4.2 The Relevance of Principles

The foregoing aims of an idealized theory directly point at the relevance of identifying its core principles.

The nature of principles Before asking about the use of core principles, it doesn't seems uncalled for to ask what "principles", as I use the term here, actually *are*.

This must remain a bit vague. What I mean to refer to are the fundamental ideas behind the theory, or the theory's main higher-level aspects and characteristics. Principles, as such, would be invoked to give an intuitive impression of the theory, or to defend its value or validity. They indicate why the theory is cogent. In this broad description, principles could possibly take the role(s) of

- a supposition, or
- a consequence, or
- a justification

of the theory. Rather than trying to make this intuition more precise, and risk cutting off useful associations, I hope that in the course of this chapter it will become clear why it's natural and helpful to pick out some aspects of the theory and give them the special status of "principles".

The use of principles Naturally, identifying one or more basic principles that ground Solomonoff's theory will give a better conceptual appreciation of the theory. But the interest of doing so goes beyond the theory itself. It's also bound to contribute to the fulfillment of the theory's potential as an ideal theory.

First of all, a clarification of philosophical issues via the theory will only benefit from a recognition of the fundamental principles that the theory is founded on.

Furthermore, rather than the technical details surrounding them, it will be the core principles of the ideal theory that serve as inspiration for theories expanding on it. This is the main justification for undertaking the project in this Chapter, and its relevance for any forthcoming discussion of successor theories.

2.2 The Candidate Core Principles

If we agree on its general interest, we can embark on the project of identifying the core principles of Solomonoff's theory. In this Section, I will first extract three candidate core principles from the main aspects of the theory (Subsection 2.2.1). These three principles, namely the principles of Completeness (Subsection 2.2.2), Simplicity (Subsection 2.2.3) and Universality (Subsection 2.2.4), I will then inspect in some more detail, laying the ground for a defence of Universality as the main principle in the next Chapters 3 and 4.

2.2.1 Identification of the Candidates

There are a number of general ideas that are often connected to the merits of the theory.

A list of names Recall the quote at the beginning of this chapter, recounting a couple of names that stand for the central ideas of SP. Hutter is more concise still:

> "All you need for universal prediction is Ockham, Epicurus, Bayes, Solomonoff, Kolmogorov, and Turing." [83, p. 45]

Let me connect the names with the ideas that they may represent:

- **Ockham**: the preference for *simplicity*;
- **Epicurus**: the *principle of multiple explanations*;
- **Bayes**: the *probabilistic* setting of confirmation theory (it seems only just to also include the name of **Carnap** here);
- **Solomonoff**: the *completeness* of the universal predictor;
- **Kolmogorov**: the identification of simplicity with *shortest description length*;
- **Turing**: the *effective* setting of computability theory.

I add to this list the name of *Shannon*, the father of information theory:

- **Shannon**: the *information-theoretic* language of bit strings.

The candidate core principles I will now distil from the previous list three candidate core principles, where all of the above names find their place.

- *Universality.*
  The first candidate main principle, that I will later argue is the single fundamental principle of theory, is that of the Universality of the theory. I share the three starting assumptions behind the mathematical setting of SP (as mentioned in

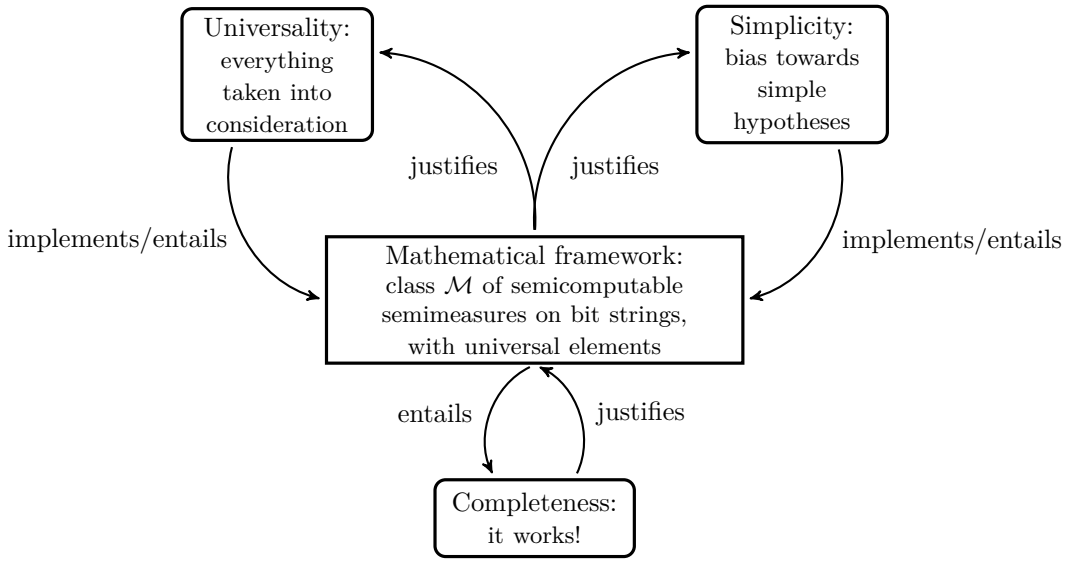

Figure 2.1: The candidate core principles and their possible relations to the mathematical framework.

Subsubsection 2.1.3.1), represented by **Shannon**, **Bayes** & **Carnap** and **Turing**, under this principle. This is because, as I will show later, their heuristic relevance is mainly to keep things as general as possible. The specific instantiation of **Epicurus**'s principle of multiple explanations should be seen as an aspect of Universality too.

- *Simplicity.*
  The second candidate main principle is that of Simplicity, often referred to as the principle of **Ockham**'s razor. The intuition behind **Kolmogorov** complexity is to give a precise quantification of simplicity, in terms of shortest description length.

- *Completeness.*
  Solomonoff's completeness result is supposed to rigorously show that the theory is valid. I promote Completeness to the third candidate main principle, with the broader denotation that "the theory works", because this is sometimes taken as the only real justification of the theory.

I have depicted the three principles and the possible relations to the mathematical framework in Figure 2.1. A more complete (but less orderly) picture would include the same or similar relationships between the mutual principles.

The next Subsections 2.2.2 (Completeness), 2.2.3 (Simplicity) and 2.2.4 (Universality) are devoted to a more detailed exposition of our three principles.

2.2.2 Completeness

Recall that Solomonoff's "completeness" of the universal predictor **M** refers to the formal results of Chapter 1, Subsection 1.4.2: convergence of the predictions of the universal predictor to the actual probabilities. It means that using **M** will almost always and very quickly lead to the best possible predictions.

2.2.2.1 Completeness as a Principle

It might seem odd to treat this completeness as a candidate core principle of the theory. I have included it here to address the more general conviction that it's enough for the theory to "work", to give appropriate results: in this case, accurate predictions. Naturally, this aspect is of crucial importance – sometimes it's taken as the only argument that matters. Solomonoff himself is very explicit about this:

> "It should be noted that the only argument that need be considered for the use of these techniques in induction is their effectiveness in getting good probability values for future events. Whether their properties are in accord with our intuitions about induction is a peripheral issue. The main question is "Do they work?" As we shall see, they *do work*." [168, p. 1]

So the fact that the theory works can be its only possible justification. According to the strictest version of this way of looking at things, it's a pointless exercise to try to defend SP by means of other principles: it suffices to refer to the completeness result. To do justice to this prominent point of view, I posit Completeness here as a core principle in its own right.

Completeness of a method However, I will be quick to dismiss it again in the next Chapter.

The reader would have noticed that an adherence to Completeness in said manner is very much related to adopting the method interpretation of SP. Certainly, the first thing that matters for a method is that it works properly. The dismantlement of Completeness in Section 3.1 of the next Chapter, aided by the more systematic perspective of the principle of Universality, then also involves, in fact builds on, the conclusive dismantlement of this method interpretation.

2.2.3 Simplicity

The principle that probably pops up most often in discussions of Solomonoff's theory is that of Simplicity.

2.2.3.1 Simplicity in Science

The preference for simplicity is a main heuristic in scientific practice and theory [3, 98]. Our familiarity with its use and obvious-looking virtues is also a reason why its exact form and application is often not made very precise; in many cases, recognizing the

"simplest" of a number of competing theories is a matter of immediate intuition, and its selection proceeds with only a nod to the principle of *Occam's razor* for justification.

The standard reading of the razor principle informs us that "entities are not to be multiplied beyond necessity". It would probably be a disappointing exercise to measure the force of this heeding by the number of scientists that are now kept from multiplying entities in their theories beyond what they deem necessary. The principle, so stated, is either too vague or too obvious to tell us anything about what simplicity *is*, and *why* we should look for it.

If we need to compare hypotheses to their simplicity, and we want to do this in a transparant and objective way, we would need some *measure* of simplicity. We would need a *definition* of one hypothesis being simpler than another. This forms the basis for a *justification* of our preference for simplicity.

Definition and justification of simplicity We can distinguish two broad types of simplicity. First, there is *syntactic* simplicity, or *elegance*, of a hypothesis: the conciseness or clarity of the hypothesis itself, say when written down. Second, there is the *ontological* simplicity, or *parsimony*: the number of things that are postulated, and their complexity.

In justifying why we should aim for simplicity in the first place, there are some further useful distinctions to be made. We may look at the simplicity principle as a *methodological* principle, that says that it's sensible, in practice, to (provisionally) prefer the simpler of two valid hypotheses to work with. One reason could be that we adopt simplicity as a goal itself, that needs no further justification: a more elegant or parsimonious hypothesis is, other things being equal, obviously more desirable to work with. Another reason could be that we believe that the simpler hypothesis is more likely to pay out: it's more likely to generalize or predict better. A stronger position than the methodological one is to take simplicity as an *epistemic* principle: it's rational to *believe* in the simpler hypothesis. In that view, simplicity serves as a pointer to the truth. (Note that these readings of Occam's razor are still different from a *metaphysical* position that the world must be simple.)

Whatever its precise form and motivation, I will refer to the adoption of a preference for simplicity by the *principle of Simplicity*.

2.2.3.2 Simplicity of Descriptions

Right from the beginning, Solomonoff stressed the connection of his theory to the principle of Simplicity (also under the label of Occam's razor), in the form of a preference for *shorter descriptions*:

> "That these kinds of models might be valid is suggested by 'Occam's razor,' one interpretation of which is that the more 'simple' or 'economical' of several hypotheses is the more likely. Turing machines are then used to explicate the concepts of 'simplicity' or 'economy'—the most 'simple' hypothesis being that with the shortest 'description.' " [163, p. 3]

This evocation of Occam's razor was taken over by virtually all authors that later wrote about Solomonoff's theory (also see [42, 176, 67], and the quote in the introduction to this Chapter):

> "Occam's razor is equivalent to choosing the shortest program that produces a given string." [32, p. 161]

> "(...) we arrive at a predictor that favors simple descriptions in essentially the same way as advocated by Occam's razor." [140, p. 55]

The short descriptions of algorithmic probability The short or simple descriptions that are mentioned in the previous quotes refer to the minimal descriptions in the definition

$$Q(\sigma) := \sum_{\tau \in T_\sigma} 2^{-|\tau|}$$

of algorithmic probability (repeated here for convenience, see (1.2) on page 29).

Clearly, the shorter the minimal descriptions τ of given string σ, the greater the algorithmic probability of σ. In the experiment of presenting random noise to universal U, the probability of arriving at a shorter description is naturally higher; and so is the probability of consequently obtaining a string with a shorter description.

The bias towards simplicity The assignment of higher probability to strings with shorter descriptions is one step. Identifying *shortness of descriptions* with *simplicity* is the next. Taken together, the simplicity bias of algorithmic probability follows.

Uncontestably, strings with short descriptions are in a precise sense quite simple. Strings with very short descriptions can be said to be highly *compressible*: their actual information content is low; they are simple. On this reading, the shortest descriptions via a universal machine are not just about syntactical simplicity – this description length truly reflects the intrinsic information content of a string. If there are short descriptions of σ, then there is some algorithm that can detect patterns in it, rendering σ relatively simple from the corresponding machine's perspective; if there are none, no machine can reduce the complexity of σ, which must mean it is truly complex. The shortest description lengths give a measure of all structure that is traceable in an effective way, which, if we take a string to represent a hypothesis, must even include such things as its parsimony, hence ontological simplicity. It seems that, in the universe of binary sequences, we are on the track of a genuine *definition* of simplicity.

(This account should look familiar from the motivation of Kolmogorov complexity on page 24. The subtle yet important difference is that Kolmogorov complexity only considers the length of the one *shortest* description – I will come back to this point later.)

All in all, the description lengths of a string appear to provide a convincing measure of its simplicity, and so algorithmic probability would show a clear preference for simple strings.

2.2.3.3 The Outlook on Simplicity

This concludes our precursory discussion of simplicity and Occam's razor, and the reason why Simplicity is regularly associated with Solomonoff's theory. On this first glance, the theory doesn't only point at a *definition* of simplicity in terms of minimal

description lengths, but also implements a direct preference for simplicity, be it as a starting *assumption* or as a *consequence* of more fundamental considerations. In both cases, the promise of the theory is an honest *justification* of a preference for simplicity; indeed, since this preference provably leads the universal predictor to the assumed actual, *true*, distribution, a justification of the *epistemic* kind: an exciting prospect.

In Section 3.3 of Chapter 3, I will review these points in more detail. I will elaborate on the exact guise of simplicity in the definition of algorithmic probability, which includes a discussion of the relation between Kolmogorov complexity and algorithmic probability as two related yet distinct measures of string complexity. In addition, I will examine the role of simplicity in the equivalent definition of the universal mixture distribution.

2.2.4 Universality

The content of the final candidate core principle, the principle of Universality, is perhaps not immediately as clear as that of the other two. We've already encountered the adjective "universal" in more than one context (specifically, the notions of universal machine, universal semimeasure, and universal prior distribution), but what is it supposed to mean as a principle?

2.2.4.1 All Things Considered

The notion of *universal element* in the class of semicomputable semimeasures (as given by Definition 1.4, page 33) refers to an element that *dominates* every other element in its class, that assigns probability values that exceed those assigned by all other elements (up to a constant, of course). Similarly, the notion of *universal machine* (Chapter 1, Subsection 1.1.3) refers to a machine that can emulate all elements of the class of Turing machines. The gist of the meaning of the technical notion of *universality* is that of *being able to simulate everything* (relative to a given class).

Universality as I wish to take it here has a somewhat more general denotation still. The way I will use it, the intuitive meaning of the principle of Universality is that *everything is taken into consideration.* In Solomonoff's theory of Prediction,

- every conceivable possibility finds a place in the formal framework, and
- no possibility is disregarded by the universal predictor.

2.2.4.2 Universality of the Setting

The first item requires a framework that is sufficiently unrestrictive to accommodate all conceivable possibilities. Specifically, a model that allows us to capture, in principle and with due amount of abstraction, every predictive problem; and that for each such problem includes every possible hypothesis, i.e., every possible prediction.

The generality of the model boils down to

- the generality of the information-theoretic language of *binary strings*, and

- the generality of the probabilistic setting of *semimeasures* on binary strings, and
- the generality of the effective setting of *semicomputable* semimeasures on binary strings.

That is, the generality of our formal environments as representations of real-world environments. I consider the starting points of a probabilistic and effective setting in an information-theoretic language as instances of Universality, because they serve to erect a framework that is maximally unrestrictive, maximally general – a universal model.

2.2.4.3 Universality of the Predictor

Solomonoff's theory also adheres to the principle of Universality because the predictor $\mathbf{M}$ is universal. This predictor can be seen to take each and every possibility into account, and, moreover, to keep all options open. Formally, it dominates every other predictor, giving rise to the important completeness properties.

All programs, and all semimeasures Definition 1.2, page 29, of algorithmic probability Q is based on a universal machine, that represents *all* machines, and on *all* minimal descriptions for a given string via that machine. Thus *all* programs that could generate the string – intuitively: *all its possible causes, all its possible explanations* – are considered.

As discussed in Chapter 1, Section 1.3, the aim of universality specifically drives the equivalent Definition 1.5, page 33, of the universal mixture distribution ξ. It gives a weighted sum of *every single* semicomputable semimeasure – again, intuitively: *every possible cause, every possible explanation.*

All possible extensions Furthermore, the probabilistic setting enables predictions that keep all options open. The domination property of the universal predictor ensures that if there is some measure that attributes a nonzero probability to a specific future (and there always will be), then the universal predictor will assign it a nonzero probability too. Hence every future will receive a (possibly vanishing) share in the total probability, no future is definitely ruled out.

Epicurus' stamp of approval Recall from Chapter 1, Subsubsection 1.2.1.2 that Li and Vitányi evoke as support for Solomonoff's theory Epicurus' principle of multiple explanations, that reads that *all hypotheses consistent with the data should be retained.* (This is taken over by Hutter.) Both of the above properties, the consideration of every explanation (program, measure) and the preservation of every hypothesis (continuation), can be said to agree with the gist of this principle. Even so, it seems a bit opportunistic to take out this single instruction from the rich context of Epicurus' philosophical system[2] and, via the above agreement, employ it as testimony for the validity of Solomonoff's theory. (Not to mention their treating in

[2] See, for instance, [2] for Epicurus' epistemology and the precise role of the principle of multiple explanations in this context.

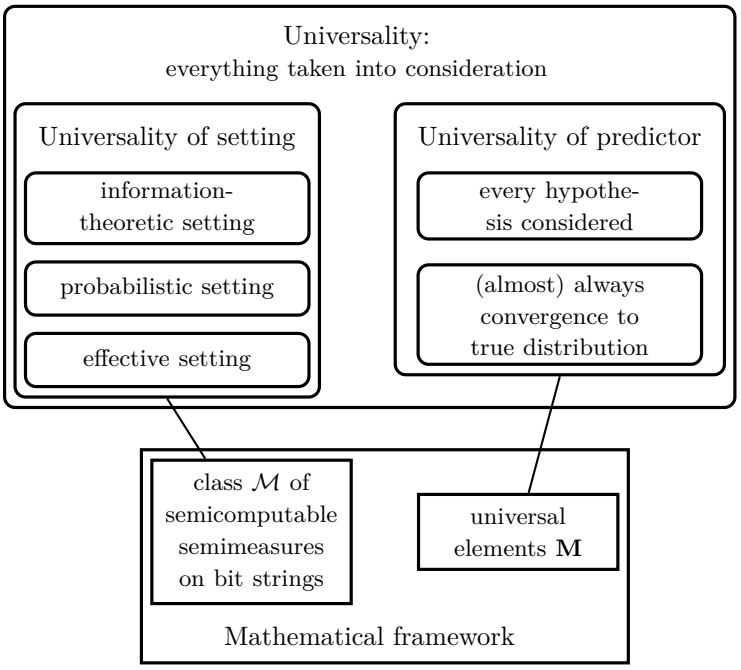

Figure 2.2: Universality.

the same breath Epicurus' principle and the historically and philosophically quite unrelated principle of indifference.) It seems more sound to say that this heuristic directive is satisfied, in Solomonoff's setting, simply as a consequence of the much more fundamental principle of Universality.

Universality as domination I mentioned the previous interpretations of the functioning of the universal predictor because they are illustrative of its universality. Formally, it really all comes down to the single property of domination of every other effective semimeasure, every other predictor. The fact that the universal predictor's assigned probabilities at least match those of any other effective semimeasure, within a constant bound, will lead it very quickly to predict just as well as any other predictor – as codified in the completeness results.

2.2.4.4 The Outlook on Universality

This subsection served to give a first impression of the bearing of Universality. I will discuss the universality of the predictor and the setting in much more detail in the next Chapters 3 and 4, respectively.

The bottom line is that in constructing a framework that is universal in the sense of keeping near full generality, incorporating only restrictions that are sufficiently mild

to hardly touch its inclusiveness but sufficiently substantive to allow us to infer non-trivial conclusions, we are in fact lead to a significant result: the existence of universal prediction methods. From Universality, we get Universality. According to this interpretation, the whole theory can be said to be an instantiation or consequence of the principle of Universality, which therefore should be seen as the one and only core principle of Solomonoff's theory of Prediction.

3 Universality of Solomonoff Prediction

We have inherited from the previous Chapter three candidate core principles of Solomonoff's theory of Prediction. In this Chapter, I will further reduce this number. In a more in-depth analysis of the other two candidate principles, I will expand on my proposed interpretation of SP, and highlight the fundamental role of the principle of Universality.

First I take a closer look at the principle of Completeness, show why it's best seen as just an aspect of Universality, and conclusively discard the method interpretation of SP in the process (Section 3.1). Next I will discuss a supposed main threat to Universality, namely the element of arbitrariness introduced by the choice of a particular universal predictor. This discussion, too, ultimately turns into a demonstration of the prominence of Universality in SP (Section 3.2). After that I examine in depth the role of the principle of Simplicity, and conclude that this principle should also be seen as subsidiary to Universality (Section 3.3).

This Chapter resolves around the Universality of the universal predictor, and presupposes the generality of the framework that it operates in. In the next Chapter 4, I turn to a defence of the Universality of this setting, of the model of SP.

3.1 The Encapsulation of Completeness

I introduced Completeness as the third and final candidate core principle. Now I will retract it again, merging the completeness results into Universality.

3.1.1 Making the Method Work

The completeness Theorem 1.9 is an essential result of Solomonoff Prediction. It shows that the single universal predictor will almost always convergence on optimal predictions, no matter what the actual measure μ is, and it shows that this convergence is extremely rapid. Certainly, without this result the theory would lose much of its interest.

But is this sufficient to show that the theory "works", as the principle of Completeness has it? Can the completeness result bear the burden of an insightful justification of the theory, and as such reasonably figure as a proper principle?

That this formal property makes the theory work is an interpretation that is attached to what is, in the first place, a technical result within the mathematical framework. We could start undermining this interpretation by pointing at the loopholes in the

convergence results. In particular, recall from Corollary 1.10, page 42, that it's *only with μ-probability 1* that the probability values predicted by $\mathbf{M}$ converge to the actual values of μ. This still leaves a class of measure 0 of μ-generated sequences on which $\mathbf{M}$ doesn't converge. In fact, Hutter and Lattimore [110] have shown during the writing of this thesis that for all universal mixtures ξ_w there are Martin-Löf random sequences on which it fails to converge. The class of Martin–Löf random sequences can be seen to contain the most typical sequences generated by the uniform distribution $\mu = \lambda$ with $\lambda(\sigma) = 2^{-|\sigma|}$; the fact that $\mathbf{M}$ is unable to recognize some of these sequences is not an insignificant limitation.

A further thing we could ask is whether we should be conclusively satisfied with just the completeness properties of the universal predictor's pointing the way towards accurate probability values – when it's perfectly possible to also take into account, for instance, matters of efficiency (like the number of computation steps in calculating these values) in the framework.

I won't go into these particular worries. I would say that completeness as definitively showing the theory to work is too strong a reading, but for a much more fundamental reason. What the completeness result asserts, at best, is really only that the specific *method* of the universal predictor works. As such, it is associated with the interpretation of SP as a method of prediction (Chapter 2, Subsection 2.1.2). An argument that Completeness is not a fundamental principle of Solomonoff's theory should therefore start with a rejection of this method interpretation of SP.

3.1.1.1 Against Method

We saw in Chapter 2, Subsection 2.1.2 that there is no hope of applying or approximating the universal method of SP in a practical way. Its virtue must be a different one – one that is barred by the method interpretation, as I will begin to argue.

Only a method My motivation for dismissing the method interpretation of SP is superficially connected to an important critique of Bayesian confirmation theory by K.T. Kelly and C. Glymour [96]. They attack Bayesian confirmation theory on precisely the point that it give nothing but a *method* of inference. In their view, the theory comes down to just one strategy for finding the "truth" (or more generally attaining a consistent theory, or accurate predictions[1]). Therefore the theory is not fit to reach for what they take to be the main goal of a philosophical theory of scientific inference: providing an explication of scientific justification. Contrary to received wisdom, Bayesian inference theory is not even the right *sort* of thing.

Their specific argument is couched in the requirements that a justification of scientific inference should accommodate, first, the intrinsic difficulty of inductive problems, and, second, the relative efficiency of the method used. Bayesian confirmation theory, so they claim, achieves neither: the probabilities of confirmation theory – probabilities that hint at actual, permanent, results, but that in reality can keep on fluctuating –

[1] In their own words: "The reference to truth (...) should not be taken too seriously. (...) let "truth" range over all such finite-evidence transcending cognitive goals" [96, p. 98].

give no clue about the method's performance on any specific problem in an absolute sense, nor in comparison with other methods.

One way of replying to Kelly and Glymour is to question the specific requirements on a theory of scientific justification that they postulate. I don't want to expose ourselves to such objections: the lesson that I wish to distill from the argument is the overarching point that *a specific strategy of inference, by itself, gives us little with respect to the aims of a (philosophical) theory of inference*, with respect to a clarification of the related conceptual issues. In fact, from that more general perspective, the first point above doesn't seem imperative: surely it would be too strict to exclude the possibility of philosophically valuable theories of inference that don't provide a measure of difficulty of inductive problems. It's only one example of an aspect that such a theory could elucidate.[2]

The second point looks more fundamental. It reflects the idea that the sole act of advancing a certain inference strategy is philosophically not very interesting, for the specific reason that the result would be just one among other possible strategies, that might be just as good or even better. But in the case of SP, doesn't the completeness result address just this concern, in showing that the universal prediction method is truly optimal? It does show a specific optimality – under the assumption that the true environment is given by an effective measure. So we could start by noting that we need to pass beyond the method because a full demonstration of the optimality of the method should also include a justification of the generality of the *model*. But let's say that we unreservedly accept completeness as showing the universal predictor to be optimal: then we still haven't escaped the gist of the above point. An optimal method in a specific sense doesn't exclude alternative methods that are just as good in the same or some other sense. Again, however, I want to take this observation as just one illustration of the more general point: namely, that a proposed method, even if it provably has desirable technical properties (like those expressed by completeness), simply needs more context to have any bearing on the broader conceptual matters. Even the specification of an optimal method of prediction is by itself not sufficient to arrive at any philosophical conclusions.

Competition: formal learning theory As an alternative Kelly and his collaborators promote *formal learning theory* (FLT) ([134, 63, 97], also called *logical reliability theory* [93] or *means-ends epistemology* [151]).

In FLT, the problem of induction (or prediction) is subdued *not* by weakening the notion of truth (as confirmation theory does by replacing it with the, in Kelly and Glymour's view, questionable notion of confirmation), but by giving up the requirement that inductive procedures should halt or "ring a bell" when they succeed. The best possible way to convergence to the truth in an inductive problem may, due to the typical intrinsic difficulty of the problem, only lead to assured success in the limit; and such a truth-finding strategy can be said to be justified if it solves the problem in this way in the best possible sense, that is, with the least maximal number of retractions of conclusions or "mind-changes". The analogy to formally unsolvable, or incomputable, problems is a powerful heuristic in this approach, and just like

[2] From this more general perspective one might also criticize Kelly and Glymour on their not taking sufficient account of the possibility of a broader *theory* conception of Bayesian confirmation (cf. Chapter 2, Subsubsection 2.1.1.1).

the familiar classification of formal problems by their computational complexity, we can build a hierarchy of empirical problems based on their intrinsic complexity, a complexity that is characterized *topologically* (according to Kelly [94], topology can best be thought of as the mathematical theory of ideal verifiability) by the best possible ways in which they can be solved.

It's interesting to note that while SP and FLT are radically opposing theories in many respects, they share the theory of computability as point of departure.

A proper theory The brute fact of the specification of an (optimal) inference strategy does little to illuminate the related conceptual issues. But what else could be the purpose of an *idealized* strategy, that withholds the promise of ever being directly implemented or closely approximated? What merit could we find in such an ethereal strategy, other than an indirect one: to provide a better understanding of the related conceptual issues, or, much related, to inspire higher-level methodologies? Indeed, these are the aims of an idealized theory that we identified in Chapter 2, Subsubsection 2.1.4.2 – and it would only be in the context of such a theory, that puts the method in a broader perspective and thence promises a handle on the much more interesting matters just *how* or *why* it is good, that these aims can be fulfilled. The conclusion must be that the theory of Solomonoff Prediction should be treated as just that, a *theory* – not as a stand-alone method. If SP is to have a wider relevance to the challenges surrounding scientific inference, this is the way to look at it. One aim of this entire Chapter is to show that we *can* look at it in this way.

3.1.2 Making the Theory Work

The dismissal of the method interpretation brings with it the rejection of Completeness as a principle.

3.1.2.1 Making an Idealized Theory Work

The principle of Completeness expresses that the theory works because of completeness. But while it certainly is an essential component of the theory, the result of completeness only states, on a charitable reading, that the universal predictor works. Considering the aims of an idealized theory, the requirements on a "working" idealized theory should be much broader: it might be said to "work" just in case it can fulfill these aims. The completeness result is conceptually way too limited to bear this burden. It doesn't make the theory work.

Since the purpose of the core principles was precisely to further the aims of the idealized theory, the completeness result shouldn't be considered a core principle. But, if anything, the principle of Universality would.

3.1.2.2 An Aspect of Universality

The completeness property of the universal predictor is a direct formal consequence of its property of domination, of formal universality. Conversely, if a desirable property

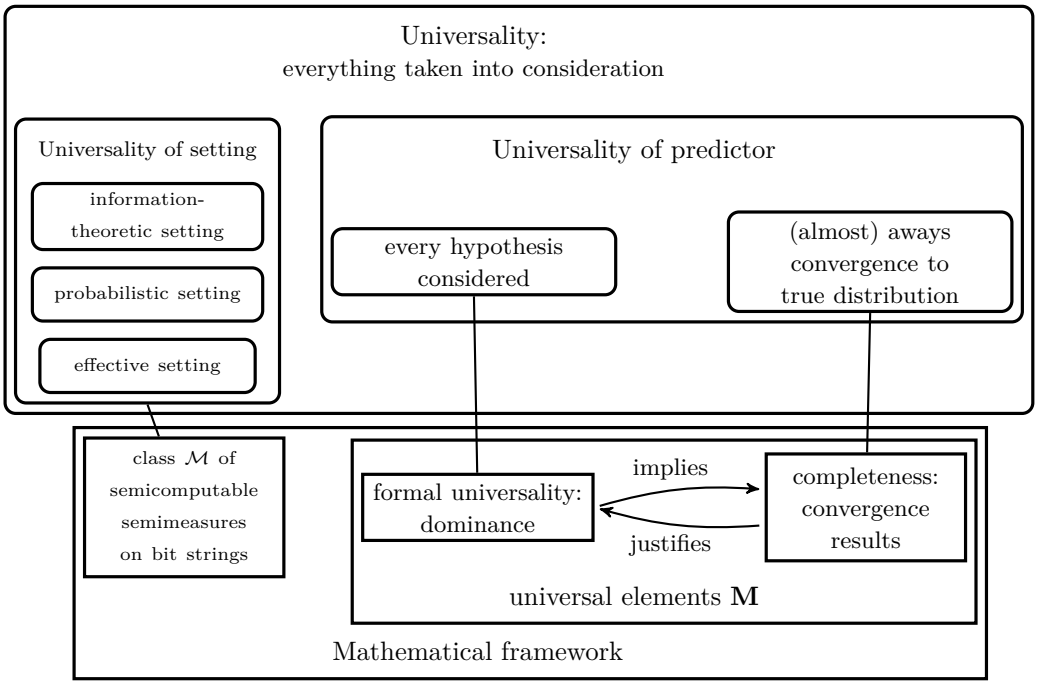

Figure 3.1: Completeness and Universality.

can be shown to arise from a starting assumption, the former can serve as a justification for the latter. In that respect, the completeness result serves as a justification of the significance of the formal property of universality.

Thus formal universality and completeness are two sides of the same coin. Indeed, completeness, rapid convergence to *any* actual measure, can rightly be seen as an additonal universality property of the universal predictor.

Formal universality is a clear instance of the principle of Universality, as part of a setting that is motivated by Universality and as the formal property of predictors that adhere to Universality. Therefore the dual property of completeness is an instance of Universality too. Even if we interpret the completeness result as an independent consequence and justification of Universality, and I think this is a valid way to look at it, completeness certainly can't be said to be a principle itself, and so it should be treated as just an aspect of this much more fundamental principle of Universality.

3.2 The Threat of Subjectivity

There is one important problem connected to the principle of Universality that merits our attention at this point. In response to this problem I will lay down more clearly what I think is the essence of SP: the existence of universal predictors in a universal

setting.

3.2.1 The Threat

In most of this Chapter, I conveniently talked about *the* predictor $\mathbf{M}$ – as I said I would do at the end of Chapter 1, Section 1.3.2. But, of course, this $\mathbf{M}$ is not uniquely defined. There are infinitely many ways of defining a universal semicomputable semimeasure along the lines of Definition 1.5 – one for each bounded weight function. Analogously, there are infinitely many ways of defining algorithmic probability as in Definition 1.2 – one for each universal monotone machine.

It's customary to bring in the Invariance Theorem 1.3 (page 30) as proof that this freedom is nothing to worry about: all possible choices are formally equivalent. The difference in probability values given by any two universal predictors is strictly within the bounds of their unique multiplicative constants.

Still, these constants could be gigantic. While this becomes increasingly inconsequential for larger sequences, for short sequences it could make all the difference. Even the Invariance Theorem can't confute the impression that the unmotivated stipulation of a particular reference weight function or universal machine brings in a strong element of arbitrariness. Accordingly, it has been recognized by the proponents of Solomonoff's theory as one of its foremost challenges (cf. [84]).

3.2.1.1 The Quest for a Unique Universal Predictor

Keeping to the definition of algorithmic probability for the moment, things would look better if it were possible to single out an objective or "natural" machine to define Q with. Some universal machines are obviously more natural than others: a machine that has a number of complex strings hard-coded (and consequently accepts very short descriptions for these complex strings) is strongly biased towards these particular strings and seems less general than machines that aren't. The intuition is that natural machines should contain as little information as possible themselves. One would make this precise in terms of description length of a program that gives this machine – but of course, that requires another natural machine to interpret it, and we're back where we came from.

Algorithmic probability of machines In a recent paper [125], M. Müller attempts to define a distribution on universal machines themselves. Using the fact that one *can* objectively quantify how hard it is for one machine to emulate another (namely, by the probability of arriving at a description for the first of the second), he exploits the intuition that more unnatural computers are harder to emulate. This he makes precise via a Markov chain of all universal machines, that are linked by their emulation probabilities; the hope is that this chain is positive recurrent, and so yields a *stationary algorithmic probability* on all machines. Given this distribution on machines, it would be straightforward to define an objective, "machine-independent", version of algorithmic probability.

Unfortunately, this stationary distribution doesn't exist. Attaching a physical interpretation to this negative result, Müller turns it into a plausibility argument that there is no way to get completely rid of machine dependence, not via any approach.

An optimal weight function If we approach the problem of subjectivity from the definition of the universal mixture distribution, the goal would be to find a single weight function that is more objective or natural than others. Arguments have been advanced that, indeed, some weight functions are preferable because they are *optimal*. Since the weight functions in question implement a bias towards simplicity, I will discuss this strategy in the next Section 3.3 – it suffices to say here that this strategy, too, doesn't look too promising.

3.2.2 Taking the Threat Away

So it seems that we'll have to accept that there may not be a way of agreeing on a unique truly objective machine or weight function. How to proceed in that case?

3.2.2.1 Acceptance

Hutter and Solomonoff represent two possible solutions: working around the problem of subjectivity, and embracing it. They can be seen to fit into a method and a model interpretation of SP, respectively, and must give way for the final solution that springs forth from viewing SP as a proper theory.

Working around subjectivity: the method The choice of a particular universal predictor only really matters for prediction of relatively short sequences, because all such predictors will eventually converge to the actual distribution. Hence, Hutter observes, the problem could be circumvented by simply conditionalizing on a very large "prior" bit sequence:

> "It is worth noting that this problem of arbitrary predictions for short sequences x is largely mitigated if we use the method of prefixing x with prior knowledge y. The more prior knowledge y that is encoded, the more effective this method becomes. Taken to the extreme we could let y represent all prior (scientific) knowledge, which is possibly all relevant knowledge. This means that for any x the string yx will be long and therefore prediction will be mostly unaffected by the choice of universal reference machine." [140, p. 68]

While this is undeniably true, one can't help feeling a bit uncomfortable with such a solution. For what does it solve? What it gives, basically, is a rather improvised way of adjusting a method that, as an idealized method, we can never employ in this form anyway. I don't think speculations along these lines are very helpful: they are a clear consequence of sticking to the method interpretation of SP, that was dismissed in Subsection 3.1.1.

Embracing subjectivity: the model Solomonoff has not only come to accept subjectivity, but has positively embraced it, stating that it is "not a *Bug* in the system" but a "*Necessary Feature*" [170, p. 600]. In essence, his argument is that one always needs *a priori* information to perform sensible prediction. And the selection of a particular universal machine, which is really the selection of a particular *computer language* to use (that includes the general definitions that express our basic knowledge), hands us the opportunity to insert this *a priori* information – indeed, it's the most general way of doing so. [168, s. 2]

In the context of artificial intelligence research, Solomonoff himself took much of a method interpretation of his theory, as a method of *learning*. This is witnessed, too, by his talk about a *system* of induction. But what he does in this argument is of a more broad and speculative nature. He takes the machines to represent the *a priori* knowledge that is always present in a learning situation: this solution to the problem of subjectivity is an additional claim about what is *modeled* by his theory, what is captured in the *model*.

This strategy thus amounts to an additional assumption, in the form of an extension of the model. Its merit is that it, on accepting it, directly hands us an answer *why* the element of subjectivity is unavoidable: there must be room for prior knowledge. Of course, the problem is to give independent rationale for adopting this supposition in the first place. Any additional assumption poses a potential threat to the generality of the model, in its ability to veraciously reflect the real world. But even if we find it harmless enough in this case (perhaps on the grounds that it only concerns the ethereal universal predictors), further arguments are required for the inevitability of what is undeniably a rather speculative interpretation of the difference between the various universal elements. Otherwise, this solution doesn't amount to much more than a recognition of the unavoidability of subjectivity and just one non-binding possibility of how to make sense of it.

Beyond subjectivity: the theory The previous two solutions weren't very satisfying because they took too narrow a view of the theory. Going beyond the levels of method and model, and taking a renewed look at the role of Universality in connection with the aims of the idealized theory, will also take us beyond the problem of subjectivity.

3.2.2.2 The Class of Universal Predictors

My proposal is that the content of Solomonoff's theory of Prediction consists in the result that in a universal framework, there exist universal predictors. In the information-theoretic, probabilistic and effective framework, *there exist* predictors that take everything into consideration, and that eventually predict as well as any other.

It doesn't matter that there is an entire class $\mathcal{U}_{\mathbf{M}}$ of such predictors. It doesn't matter that among them, none can be said to have a privileged status. We won't ever actually have to choose one such predictor, let alone use it. The important thing is that they can be shown to exist.

Indeed, if no member of this class of predictors has a privileged status, this is only justification for treating the whole class as one predictor: the universal prior

distribution $\mathbf{M}$, *the* universal predictor $\mathbf{M}$. I'm not saying that all universal predictors are equal: in spite of the Invariance Theorem, they're not. But they all share the same property of universality, and for the purposes of the idealized theory, the existence of predictors with this property is the relevant thing.

As soon as we accept this, the threat of subjectivity evaporates. It's simply not relevant. Insisting on the selection of one particular predictor presupposes that we would actually go and employ it for prediction. We wouldn't because we can't: SP doesn't give practically applicable methods. It makes no sense to descend below the level of the class of universal predictors. The prime thing we get out of the theory is that this class exists.

One more cheer for Universality The existence of a universal predictor in a universal setting is the result that defines the theory, and that may further its aims as an idealized theory. This is the result that can, perhaps, provide an entrance to the philosophical problem of prediction. This is also the result that can, hopefully, inspire higher-level theories. And the principle that constitutes this result is the principle of Universality.

3.3 The Questionable Role of Simplicity

I hope to have made plausible by now that the fundamental starting point of Solomonoff Prediction is the principle of Universality. There is yet one major obstacle to declaring Universality the sole principle of SP: the principle of Simplicity. In their emphasis on the role of simplicity in Solomonoff's theory, most authors have us believe that Universality must at least share the honour with Simplicity.

Certainly, even if I'm granted that the main import of the theory is the existence of a class of universal predictors in a universal setting, following only from the principle of Universality, it's still highly interesting, for both of the aims of the idealized theory, to see if the predictors of this class have other characteristics in common. A bias towards simplicity, for instance.

Recall that the elements of the class $\mathcal{U}_{\mathbf{M}}$ of universal predictors can be defined both as algorithmic probability distributions and as universal mixture distributions. In Subsection 3.3.1 I will take a closer look at the role of simplicity in the definition of algorithmic probability, and in Subsection 3.3.2 I will shift attention to the definition of the universal mixture distribution. I conclude in Subsection 3.3.3.

3.3.1 The Short Descriptions of Algorithmic Probability

Consider again the definition (1.2), page 29, of algorithmic probability. As discussed in Chapter 2, Section 2.2.3, it seems to reflect an unambiguous bias towards shorter descriptions, so simpler strings. Let us now consider more closely the exact guise of simplicity in this definition.

3.3.1.1 The Definition of Algorithmic Probability

The link of simplicity with algorithmic probability is established in a two-step process. First, there is the association, inspired by information theory, of probability with description length. Second, there is the identification of description lengths and simplicity.

Simplicity from algorithmic probability Does the bias towards simplicity figure as an *assumption* or as a *consequence* of the definition of algorithmic probability? This depends on whether or not there is more to the first step than the blunt stipulation that shorter is more probable. As a way of denying this, in the hope of isolating a more fundamental starting point, one could argue that this assumption is not visible in the set-up of the bit-feeding experiment that defines algorithmic probability. Rather, the experiment starts from Universality it's a universal machine that's in the centre. This scenario has a certain intuitive appeal, but one could still question its status as starting point by inquiring about the reason for specifically feeding the machine random noise.

Another, only slightly different way of looking at the experiment is that the algorithmic probability distribution results from a transformation, through a universal monotone machine, of the uniform distribution $\lambda(\sigma) = 2^{-|\sigma|}$. Then what is the justification for the choice of this distribution? Since the choice for the uniform distribution comes down to the decision to hand out equal probability to all descriptions of the same length, this is where one might appeal to the principle of indifference (Chapter 1, Subsubsection 1.2.1.2).

Now the question is what this appeal amounts to. Do we mean to presuppose the validity of the principle as formulated by Laplace or Keynes, and then ground the validity of our choice for uniform λ in the bit string setting as a special case of the general principle? In that case, the credibility of this move is exposed to the full amount of serious problems concerning the principle of indifference (page 23). Alternatively, we formulate a more specific version of the principle that applies in our setting (a quite specific version, that partitions the class of all strings with respect to their lengths before evenly distributing probability). But clearly we don't gain anything by assuming the validity of a principle that asserts just what we set out to justify. Hence the evocation of the principle of indifference doesn't hand us a more fundamental starting point, and we are still stuck with the task of justifying the choice for the uniform distribution. Why start here – isn't it odd to presuppose a uniform prior distribution, on the silent ground that this is the unique most objective one, towards defining a universally objective prior distribution?

In the end, description length is the only thing that determines a string's uniform probability. It seems there is little we can do but to take the explicit association of probability with description lengths as the starting point. In which case simplicity is an *assumption* in the definition – at least, if we are prepared to equate simplicity with description lengths.

Simplicity in algorithmic probability There is one potential downside, though, of identifying simplicity with description lengths in the manner of algorithmic probabil-

ity. This has to do with the possibility that a string with only long minimal descriptions still receives a high probability because of the sheer *number* of its minimal descriptions. Even though shorter strings are arrived at earlier in the experiment setting, longer descriptions are certainly not neglected. A string that only has very lengthy descriptions (and so should be considered quite complex following the intuition that shortness of description signifies simplicity), but a great many of them, will still be assigned a relatively high algorithmic probability – or so it seems. After all, it might very well be the case that any string that has sufficiently many long descriptions to make this example work must also have a short description. This issue boils down to the question whether algorithmic probability is (by approximation) equal to the probability given by the single shortest description only. The measure of complexity by *shortest* description length is a familiar one: this is the Kolmogorov complexity.

3.3.1.2 The Very Shortest Description

The idea of equating simplicity with compressibility, that was first formalized in Solomonoff's definition of algorithmic probability, finds an obvious sequel in the notion of Kolmogorov complexity (see page 24). Kolmogorov complexity is generally interpreted as providing a universally objective definition of the complexity, *ergo* simplicity, of binary strings.

To repeat, whereas algorithmic probability takes into account *all* (minimal) descriptions, the Kolmogorov complexity of a string is defined as the length of its single *shortest* description. It has been argued, nonetheless, that Kolmogorov's definition is roughly the same as a definition that takes into account *all* descriptions. This is interesting, because it would further support the claim that if we accept Kolmogorov complexity as the appropriate measure of simplicity, then algorithmic probability, implementing an equivalent measure, truly incorporates a fundamental preference for simplicity.

Kolomogorov complexity as an objective measure of simplicity But why accept Kolgomorov complexity as the proper objective measure of simplicity in the first place? Writings in the fields of mathematics and theoretical computer science that treat of the subject tend to take for granted that the notion is important, exactly for the reason that it is supposed to give an intuitively convincing definition of complexity of objects. Disappointingly, there is hardly any critical discussion of this issue in the philosophical literature. A critique of Kolmogorov complexity as a measure of *information content* is given by P. Raatikainen in [137], but he grants that Kolmogorov complexity as an explication of complexity "agrees quite well with common sense". The only thing he questions is the strategy of specifying an object by means of algorithms, noting that "it is one thing to specify an object, and another thing to give instructions sufficient for finding one – not to mention for constructing one" [137, p. 13].

Even if we grant that shortest description length is an authoritative measure of simplicity, then the objectivity of the notion of Kolmogorov complexity is still jeopardized by the familiar subjectivity involved in the choice of a particular universal machine.

Again, the bounds of the Invariance Theorem of page 25 allow for arbitrarily large constant differences.

To get a better picture of how much room the definition leaves us, suppose that we think that the measure K_U specified via universal U accurately reflects the complexity of binary strings. Now for any number n there exists a universal machine U' that for the i-th longest U-description of less than n bits returns the i-th most complex string τ with $K_U(\tau) > n$. The result is that the strings σ that are most simple according to K_U (specifically, $K_U(\sigma) < n$) have at least complexity $K_{U'}(\sigma) > n$ according to U', while a similar number of strings τ that are complex to U ($K_U(\tau) > n$) are the most simple to U': up to an arbitrarily large point, one can turn complex into simple and simple into complex. Note that the incomputability of K_M makes it impossible to *effectively construct* such a machine from a description of U, but, for the simple reason that only *finite* information is required, we can be sure that U' always *exists* (simply imagine that all relevant strings are hard-coded in this machine).

Of course, this is just a rephrasing of what is already given by the bounds of the Invariance Theorem. It's an explication of what we can do within a constant bound. One can probably come up with more imaginative examples of this kind, but the constant bounds impose a strict limit on what we can do. It wouldn't be possible to indefinitely repeat a procedure of swapping short and long: even if some such procedure were effective (which means it would have to involve approximations to complexities), the corresponding universal machine would no longer be additively optimal. One could, moreover, object that the additively optimal machines we create in such examples are no longer "natural". But again there doesn't appear to be a natural way of isolating a subclass of such machines to come to a more restricted definition of Kolmogorov complexity. And the freedom that the original definition leaves us to derive intuitively very diverging measures must cast at least some doubt on the status of Kolmogorov complexity as a completely satisfying explication of simplicity.

Kolmogorov complexity and algorithmic randomness Some independent support for the objectivity of Kolmogorov complexity might be drawn from the field of algorithmic randomness, a branch of algorithmic information theory that has found a revival in the last decade (see the textbooks [127, 46]; also see page 91). Kolmogorov complexity provided a way to make the intuition precise that random sequences should be highly incompressible: a sequence is random if all its initial segments have maximal Kolmogorov complexity. A relevant fact, codified in Schnorr's Theorem [150], is that the main alternative formalizations, via the theory of Martin-Löf and the theory of martingales, isolate the same class of random sequences. From three intuitively very different angles, respectively highlighting the properties of incompressibility, statistical regularity and unpredictability, we obtain the exact same characterization of randomness – a situation reminiscent of the way the Church-Turing Thesis is supported by the equivalence of many different approaches. Indeed, the supposition that the *Martin-Löf random sequences* are exactly the intuitively random ones is called the *Martin-Löf-Chaitin Thesis* (MLCT) in [43] and defended by analogy to the Church-Turing Thesis. However, the former thesis appears less solid than the CTT. The support from confluence of the CTT rests on a great number of formalisms; in the case of the MLCT there are only three. Moreover, there are alternatives to Martin-Löf's characterization of randomness that have nice properties of their own. Martin-Löf randomness is based on *semicomputable* statistical tests, but

one could opt for weaker or stronger versions: computable tests isolate the *Schnorr random* sequences, tests that can use the halting set as an oracle result in the *2-random* sequences. Much the same holds for the choice of martingales. (See [127].) We are forced to the conclusion that this proposed defence of the objectivity of Kolmogorov complexity is far from conclusive.

Algorithmic probability and Kolmogorov complexity For the sake of discussion, let's suppose that we embrace Kolmogorov complexity as the proper measure of simplicity of binary strings. The question is whether the definition of algorithmic probability is in some sense equivalent.

Let me be more precise. On the one hand, there is the Kolmogorov complexity of given string σ, based on its single shortest description. Recall that in the current setting we are working with monotone machines, that give rise to the monotone Kolmogorov complexity

$$Km(\sigma) = \min\{|\tau| : U(\tau) \preccurlyeq \sigma\}.$$

The corresponding probability distribution, directly linking description length to probability, would be defined as

$$R(\sigma) := 2^{-Km(\sigma)}.$$

Note that $Km(\sigma) = -\log R(\sigma)$.

On the other hand, there is the algorithmic probability Q. Again: the more complex, the more improbable; but now all descriptions are considered. If we translate this distribution into a measure of complexity, we arrive at the definition

$$KM(\sigma) := -\log Q(\sigma).$$

Now the argument put forward by Hutter (see [81]) is that, informally,

$$Q(\sigma) \approx R(\sigma)$$

(equivalently, $Km(\sigma) \approx KM(\sigma)$). On a first reading, this would mean that a string's algorithmic probability is a direct measure of its simplicity. It would also mean that if a string has a high algorithmic probability due to a great number of descriptions, it must have a very short description as well. Ultimately, it would give further credence to algorithmic probability as implementing a clear preference for simplicity.

Hutter asserts that, since $Km(\sigma) \approx KM(\sigma)$, there is a clear bias towards simplicity in the definition of algorithmic probability. In his own words, "From $M(x) \approx 2^{-K(x)}$ we see that M assigns high probability to simple strings (Occam)." [81, p. 47] "This can be seen as both an intuitively appealing characteristic of $M(x)$ and as a fundamental justification for Occam's razor." [140].[3]

But how approximate are the two?

[3]Hutter denotes Q by M.

The Coding Theorem in the discrete setting There is, in fact, a precise equivalence between what corresponds to Q and R in a somewhat simplified, *discrete*, setting.

In the definition of (semi)measures on page 31, that are more properly called *continuous* (semi)measures, I set out from cylinder sets, and derived from those distributions over the corresponding strings themselves. Of course, we can skip this step, and directly define a function $\mu : 2^{<\omega} \to \mathbb{R}$ that attributes a real value to each finite string. While we want to keep to the natural restriction that $\sum_\sigma \mu(\sigma) \leq 1$, we would lose the property of subadditivity. In effect, we no longer care whether one string is an initial segment of another: the probability of the latter doesn't have to stay below that of the former. These functions constitute the class of *discrete* semicomputable semimeasures. One can construct a universal element as a universal mixture distribution, denoted ξ^{discr}, in the same way as in the continuous setting.

The discrete universal mixture distribution appears to be asymptotically equivalent to the version of algorithmic probability that is based on prefix machines:

$$Q^{\text{prefix}}(\sigma) = \sum_{U(\tau):=\sigma} 2^{-|\tau|},$$

with U a reference universal prefix machine. Just like discrete semimeasures do away with the intuitively valuable property of subadditivity, prefix machines lack the "dynanism" of monotone machines (they can't deal with potentially infinite strings; descriptions don't involve extensions), making them intuitively less suitable.

The probability distribution based on the single shortest prefix description, the length of which is the prefix Kolmogorov complexity K, is given by

$$R^K(\sigma) = 2^{-K(\sigma)}.$$

The asymptotic equivalence of the discrete universal mixture and the discrete algorithmic probability again motivates treating the two as representing a single *discrete universal prior distribution* $\mathbf{m}$. Now the full asymptotic equivalence of all these notions is given by the following central result.

Theorem 3.1 (Coding Theorem [115])**.**

$$\mathbf{m}(\sigma) =^\times R^K(\sigma).$$

Equivalently, $-\log \mathbf{m}(\sigma) =^+ K(\sigma)$. So not only are the universal mixture ξ^{discr} and the prefix algorithmic probability Q^{prefix} equivalent (analogous to the continuous case): both the corresponding complexity measures are equivalent to the prefix Kolmogorov complexity. Only the shortest code counts; hence Li and Vitányi's choice for the name *Coding Theorem*. The result is accompanied in their textbook by words that bring to memory the argument of confluence in support of the Church-Turing Thesis:

> "In mathematics the fact that quite different formalizations of concepts turn out to be equivalent is often interpreted as saying that the captured notion has an inherent relevance that transcends the realm of pure mathematical abstraction." [119, p. 273]

A Coding Theorem in the continuous setting: dashed hope Levin [114] conjectured that an analogue to the Coding Theorem would also hold for the monotone Kolmogorov complexity:

$$Km(\sigma) =^+ KM(\sigma).$$

By definition (since the set of *all* descriptions includes the shortest one), it is the case that $KM(\sigma) \leq Km(\sigma)$. For the converse inequality, P. Gács [53] proved the upper bound $Km(\sigma) \leq^+ KM(\sigma) + K(|\sigma|)$.

However, in the same paper he showed that there is also a strict *lower* bound. Namely, for infinitely many σ, it holds that $Km(\sigma) > KM(\sigma) + A^{-1}(|\sigma|)$, where A^{-1} is a version of the inverse Ackermann function. Although this function is incredibly slow-growing, and so apparently doesn't drive too much of a gap between the two, it still means that the difference can grow without bounds:

$$\sup_{\sigma} |KM(\sigma) - Km(\sigma)| = \infty.$$

Very recently, A.R. Day ([39], also see [55]) inferred a much stronger lower bound. For any real constant $c < 1$, there are infinitely many σ such that

$$Km(\sigma) > KM(\sigma) + c \log\log |\sigma| \tag{3.1}$$

These results entail that, contra Levin's conjecture, the monotone Kolmogorov complexity and the negative logarithm of the algorithmic probability can be quite different.

Another blow Hutter [80, 82] himself added to this the sobering result that the conditional probabilities given by Q and R can grow farther apart still. They even differ to the extent that in nondeterministic environments the convergence and optimality results of Chapter 1, Section 1.4.2 no longer hold for prediction with R, making R a worse predictor.

These outcomes together don't bode well for the strategy of showing the preference for simplicity of algorithmic probability by arguing that it's not different from Kolmogorov complexity.

3.3.1.3 Conclusion

Algorithmic probability starts from a clear identification of probability and description lengths. Less clear is the next step of equating description lengths with simplicity. Only on acceptance of this identification can we call simplicity an actual *assumption* of algorithmic probability.

We saw that the strategy of showing algorithmic probability to be equivalent to the simplicity measure given by Kolmogorov complexity doesn't work because of the failure of the Coding Theorem and the different predictive properties of the analogue definition in terms of the latter. Indeed, this last fact, that Q does give rise to the full completeness result and so has much better predictive properties than the related definition in terms of monotone Kolmogorov complexity (and the same goes for *prefix* Kolmogorov complexity, that results in even worse prediction [80, 81]), is a main reason why we should stick with this definition. But it also suggests that we have reason to actually consider algorithmic probability the better candidate as a specification of simplicity. If we can put aside the worry that taking account of all descriptions is perhaps intuitively less suitable than the shortest description only, then maybe

we should rather take algorithmic probability's sum of all descriptions as the proper *definition* of simplicity of binary strings. After all, we also saw that Kolmogorov complexity as a definition of simplicity is not unproblematic anyway.

However, algorithmic probability is very similar to Kolmogorov complexity in that it is vulnerable to the same examples of driving an intuitive wedge between simple and complex descriptions by constructing malicious universal machines. This again calls into question the credibility of stipulating an identification of simplicity and description lengths. Unfortunately, all of these observations must remain a bit inconclusive, as they resolve for large part around the fuzzy matter of the intuitive extent of the freedom given by small or constant bounds – a theme that we take with us to the next Subsection.

3.3.2 The Weights of the Universal Mixture Distribution

So far, I have only looked at the definition of the algorithmic probability Q. What about the equivalent Definition 1.5 of the universal mixture distribution ξ as a weighted mean of all semicomputable semimeasures?

This definition, too, has been related to a preference for simplicity[4]:

> "Then *Solomonoff's inductive formula* $\mathbf{M}(y|x)$ (...) can be viewed as a mathematical form of Occam's razor (...)." [119, p. 358]

How so?

3.3.2.1 Kolmogorov Complexity as Weight Function

Recall that the definition of the universal mixture distribution made use of a lower semicomputable weight function w, that was only required to satisfy $\sum_i w(\mu_i) \leq 1$ and $w(\mu_i) > 0$ for all $i \in \mathbb{N}$, where $\{\mu_i\}_i$ is a computable enumeration of all members of $\mathcal{M}$.

Now consider the weight function $w(\mu_i) = 2^{-K(\mu_i)}$. (I will come to the definition of the Kolmogorov complexity of an effective semimeasure in a moment.) This function is lower semicomputable because K is. Since any set of shortest prefix descriptions must be a prefix-free set, Kraft's Inequality (Theorem 1.1, page 26), tells us that $\sum_i 2^{-K(\mu_i)} \leq 1$ also holds true. So nothing keeps us from defining

$$\mathbf{M}(\sigma) := \xi_w(\sigma) = \sum_i 2^{-K(\mu_i)} \mu_i(\sigma). \tag{3.2}$$

This is the route that Li and Vitányi take. It very much resembles the typical Bayesian strategy of defining a prior distribution that is explicitly aimed at giving high probability to simple hypotheses (e.g., [92]). Clearly, the bare possibility of this strategy falls short of showing that we are committed to an assumption of simplicity. It doesn't exclude the existence of wildly different weight functions, that don't reflect a

[4] Li and Vitányi employ "$\mathbf{M}$" to refer to the universal mixture distribution with a specific weight function, as I will explain next.

bias towards simplicity. The only thing that it shows directly is that it's possible to *accommodate* a simplicity bias in the theory (cf. [96]) – at least, if we accept Kolmogorov complexity as the proper measure of simplicity.

Kolomogorov complexity and simplicity of semimeasures In the previous Section, I already expressed some doubts about the definition of Kolmogorov complexity as giving an objective specification of simplicity. We take these doubts with us when we use Kolmogorov complexity to measure the simplicity of *semimeasures* rather than descriptions, finite bit strings. In fact, things might seem a bit more problematic still.

Strictly speaking, it makes no sense to speak of the Kolmogorov complexity of incomputable objects, for the simple reason that there are no effective descriptions of any such object – let alone a shortest one. Yet here we make reference to the Kolmogorov complexity of *semicomputable* semimeasures in $\mathcal{M}$. The solution to this riddle is that Li and Vitányi define the Kolmogorov complexity $K(\mu_i)$ as the complexity $K(i)$ of integer i. The Kolmogorov complexity of an integer i, in turn, is defined as the Kolmogorov complexity of the i-th finite string in some standard enumeration (e.g., $0, 1, 00, 01, 10, 11, \dots$). This might seem blatantly arbitrary, but the crux is that $\{\mu_i\}_i$ is an *effective* enumeration. The exact order is not essential because every other effective enumeration can be recovered by a single computable transformation, which translates into a single additive constant in the definition of K. Still, even if the description of number n is sufficient information to effectively *approximate* the value for any input of the n-th semimeasure in a given effective enumeration $\{\mu_i\}_i$, it remains a bit of a stretch to then identify the descriptive complexity of the semimeasure μ as a complete object with the descriptive complexity of n – a concern that is added to the already existing concerns with consequently identifying shortest description length with simplicity.

3.3.2.2 The Relation to Other Weight Functions

Let's pretend that we are satisfied with Kolmogorov complexity as our measure of simplicity of semimeasures. Then the fact remains that the mere existence of one weight function employing simplicity as in (3.2) is not too interesting, unless it could be taken to show that a bias towards simplicity is in fact some more pervasive component of the theory. This might be the case if every universal predictor inherits such a bias, conceivably because every weight function is in some sense close to the above one. Alternatively, if a universal semimeasure with such a bias is in some respect *optimal*.

Optimal weight functions This strategy, too, is taken up by Hutter [79, 81, s. 5.3, s. 3.6.4]. He records as a theorem that the weight function $w(\mu_i) := 2^{-K(\mu_i)}$ is optimal in the sense that it leads to the smallest loss bounds within an additive constant. Namely, for weight functions v with small Kolmogorov complexity, from the fact that [119, Corollary 4.3.1]

$$K(\sigma) \leq^{+} -\log v(\sigma) + K(v) \tag{3.3}$$

for discrete semicomputable semimeasures v (that fits the definition of weight function), by identifying σ with (the integer representation for) semimeasure μ_i we get (by sharing small $K(v)$ under the additive constant)

$$-\ln w(\mu) \leq^{+} -\ln v(\mu)$$

for all semimeasures. Thus, the bound derived in the convergence Theorem 1.9 for the universal predictor defined with w is not much larger than that of any other universal predictor – weight function w is optimal, Hutter concludes.

Note, again, the reliance on feasible constant bounds. The first inequality (3.3) above already involves an additive constant; the term $K(v)$ brings in another. Hutter tries to keep the latter constant in line by stipulating that v have low Kolmogorov complexity, but this limitation has no motivation that wouldn't make the argument circular. Worse, the argument goes through for many other weight functions in place of $w(\mu_i) = 2^{-K(\mu_i)}$, even ones that differ more than a constant factor [81, exc. 3.7]. All in all, this argument falls short of bearing the weight of the conclusion.[5]

Harking back to the conclusion of the previous Section that it's most fruitful to interpret the class of universal predictors as a single predictor, looking for optimal predictors within this class is misguided from the start. Additional properties of the universal predictors, including a bias towards simplicity, are interesting only if they hold for the complete class. Failed attempts at identifying a subclass of optimal universal predictors only supports the validity of this view.

All weight functions The remaining strategy is to make plausible that *every* universal mixture distribution in fact shares a bias towards simplicity. This comes down to an argument that every weight function implements such a bias.

A strategy of this kind has to come to terms with definitions of weight functions that appear to deny any part for simplicity. For instance, $v(\mu_i) := 2^{-i}$ is obviously a valid weight function. It results in a mixture distribution

$$\xi_v(\sigma) = \sum_i 2^{-i}\mu_i(\sigma) \tag{3.4}$$

that is also universal, but that flagrantly lacks any intuitive link to simplicity of semimeasures.

The reply would have to be that such weight functions have a (well-hidden) preference for simplicity nonetheless. Namely, they are all very close to the distribution ξ_w of (3.2) that does implement this preference. If successful, this strategy might be taken to uncover the simplicity bias as an unexpected *consequence* of a definition that didn't explicitly include it.

[5] Hutter makes the additional reservation that it's also "more of a bootstrap argument, since we implicitly used Occam's razor to justify the restriction to enumerable [i.e., semicomputable] semimeasures" [81, p. 103]. However, there is no trace of such an argument in the section he refers to. He actually motivates the restriction to the particular class $\mathcal{M}$ in [81, s. 2.4.3] by the fact that it is the largest class of (semi)measures that still contains universal elements that are effective in the weakest sense of semicomputability.

Apparently the only candidate to explicate this proximity is the property of formal universality, that implies that all universal predictors dominate each other. In particular, one can straightforwardly derive the bound $\xi_v(\sigma) \leq 2^{K(\xi_v)}\xi_w(\sigma)$. The universal mixture resulting from any weight function v never deviates more than a constant $2^{K(\xi_v)}$ from the simplicity-biased mixture ξ_w. Once again, the argument fully rests on our accepting a constant multiplicative bound as "close" – notwithstanding the obvious examples that test our intuition to the limit.

It's fair to say that neither strategy – demonstrating simplicity in all universal mixtures by showing their proximity to ξ_w, nor confuting this by counterexamples – brings the matter to a convincing conclusion. Perhaps the most expedient move is to bring the definition of algorithmic probability back in the picture, and point at the fact (Theorem 1.7, page 36) that every universal mixture ξ_v is equal to the algorithmic probability distribution Q_U via some U. Unlike the majority of weight functions, this definition in terms of description lengths has – with all the problems when we look at it in more detail – at least some immediate intuitive connection to simplicity.

3.3.3 Conclusion

Establishing a simplicity bias in the definition of the universal mixture distribution rather than algorithmic probability appears to run into many additional difficulties. Next to having to set off from a more dubious definition of Kolmogorov complexity of semimeasures, we need to expand this supposed simplicity bias in just one specific subclass of mixture distributions to a claim about the complete class $\mathcal{U}_\xi$. In the end, our best chance of demonstrating a simplicity bias in the definition of universal mixture seems to be to turn again (via the coincidence of classes $\mathcal{U}_Q$ and $\mathcal{U}_\xi$) to the definition of algorithmic probability. That is, we reduce the problem of showing a simplicity bias in universal predictors in $\mathcal{U}_\mathbf{M}$ to the problem of demonstrating this simplicity bias in the definition of Q – in which case we can pick up the thread where we left it in the Conclusion 3.3.1.3 of the first Subsection.

The problematic part of arguing for a simplicity bias in algorithmic probability is the identification of simplicity with this specific formalization in terms of description lengths. We saw that, even if we accept the link between simplicity and description length at an intuitive level, such a stipulation of a definition of simplicity is threatened by an inherent non-uniqueness, both *within* this definition and *inbetween* this definition and other superficially similar definitions. The objectivity of any of the discussed potential definitions of simplicity suffers from the fact that they only pin things down within arbitrarily large constant bounds. And the objectivity of any one specific definition further suffers from just the fact that there are multiple competing definitions, that can mutually differ by even more than constant bounds (the prime example being Km and $KM = -\log Q$). It raises the question how much formalizations are allowed to differ if we still want to call them "unique" or "equivalent" – which is, indeed, the recurring motif in all of the discussions of the current Section.

If the debate is about whether the definition ξ should be considered unique, or whether there is at least an approximate equality $Km \approx KM$, then we have to somehow decide when to still count something as "the same" or "close enough". Should we

decide that asymptotic equivalence *is* equivalence, and are we first allowed to say that Km and KM are certainly *not* approximate anymore as soon as the precise numerical difference surpasses the bound (3.1)? In statistical practice, a log log bound is generally considered negligible. But should that count in this fundamental setting? In statistical practice, constant bounds are also no problem because constants are naturally small. Again, where to draw the line? Decisive intuition is lacking.

In my view, this lack of solid ground should warn against strong positive conclusions about simplicity in Solomonoff's theory. The opposing standpoint would be that I make too much of these numerical variances, and that it's all just a matter of accepting the description lengths of algorithmic probability as our definition of simplicity. Perhaps this issue can never be resolved to everyone's satisfaction, and I'm ready to admit that none of the considerations of this Section are truly fatal to the project of locating a simplicity bias in SP. However, at the end of the day, the burden of proof is on those who want to establish the positive result that the universal predictor justifies, or at least implements, a bias towards simplicity – not on those who aren't convinced yet. Virtually all authors just take Simplicity for granted, assuming that it's obvious from the definition. But as I hope to have shown here, there is plenty of room for doubt.

The subsidiary role to Universality In my proposed interpretation of Solomonoff Prediction, the core principle is Universality. The obligation of the universal predictors to the principle of Universality is clear and unproblematic – which, as matters stand, can't be said of Simplicity. More fundamentally, the idea of Universality conceptually unifies Solomonoff's theory. Even if the doubts I have worded in this Section could be taken away, even if I were forced to admit that all universal predictors in $\mathcal{U}_{\mathrm{M}}$ exhibit an obvious simplicity bias after all, then still Simplicity is a conceptual consequence (intriguing as it would be) of the overarching principle of Universality. Even if SP would in fact provide a justification of Occam's razor, then still I would say that Simplicity is subsidiary to Universality.

The current Chapter concerned the role of Universality from the perspective of the universal predictors. The next Chapter 4 is devoted to the universality of the setting, and complements the argument for Universality as unifying principle of Solomonoff Prediction.

4 Universality of the Model

In the previous Chapter, the focus was on the Universality of the universal predictor. It has been helpful, in the analysis of the core principles associated with the universal prediction method, to be able to treat the framework of SP as a sort of isolated whole; for that reason I simply went along with the theory's commitment to the information-theoretic, probabilistic and effective setting of bit sequence prediction.

In this Chapter, we tread outside the model and look at its connection to the wider world: we finally arrive at a discussion of the Universality of the model itself. I argue that the model is universal in the sense of being maximally general and unrestrictive. The goal is to make plausible that the model could, in principle, accommodate any conceivable predictive problem. A demonstration of the universality of the model completes the argument that Universality is the core principle of Solomonoff's theory of Prediction.

Section 4.1 treats of the unrestrictiveness of the information-theoretic language of binary strings, Section 4.2 of the unrestrictiveness of the probabilistic environments, Section 4.3 of the unrestrictiveness of the effective environments. The discussion of the last aspect is a bit inconclusive; but things start to look much better if we interpret the class $\mathcal{M}$ as consisting of predictors rather than environments, as proposed in Section 4.4.

4.1 Encodings

Solomonoff Prediction addresses the continuation of a string of bits that is known to be governed by some effective probability measure. The current problem is how we can fit prediction in general into this straitjacket.

Any predictive problem can be put in the following form. *Given*: a series or collection of past data. *Required*: (a probability distribution on) the expected series or collection of data in (a part of) the future. What concerns us: can every such problem in the wider physical world be represented as a formal past (a finite binary sequence) plus an unknown formal environment (a computable measure) that governs the possible formal futures (continuations of the formal past)?

As a start, can the relevant data itself always be represented as (parts of) binary sequences?

4.1.1 Data and Binary Sequences

The circumstance that all data from the physical world relevant to a predictive problem is to be represented as binary strings has some affinity with the metaphysical position

that the fundamental substratum of the physical world *is* digital information.

Digital physics *Digital physics* (later also *digital philosophy*) is the name that E. Fredkin gave to the direction of research that follows from the assumption that the physical universe is ultimately made up of (digital) information [51, 52]. His own early ideas influenced a number of other prominent physicists, including J.A. Wheeler, who coined the phrase "it from bit". In his words: "Otherwise put, every 'it' – every particle, every field of force, even the space-time continuum itself – derives its function, its meaning, its very existence entirely – even if in some contexts indirectly – from the apparatus-elicited answers to yes-or-no questions, binary choices, bits." [184] Thus the ultimate constituents of the universe are brute differences between states.

Many authors embed this view in an ontic *pancomputationalism*, a belief that the universe itself is a computing system. This belief was first put forward by computer pioneer K. Zuse in his *Rechnender Raum* ([190], 1969); subsequent authors in this tradition regularly employ a giant *cellular automaton* as a model of the universe (S. Wolfram's controversial *A New Kind of Science* ([187]) being a famous example).

A main implication of digital physics is that at the fundamental level, the world is not continuous but *discrete*. This is at odds with mainstream physics [D. Dieks, personal communication]. While the standard adherence to a space-time continuum is by itself no conclusive ground to reject the discrete view of nature (and the picture of quantum mechanics does give rise to broad speculation in the direction of a fundamental discreteness), there is a conspicuous lack of positive evidence that speaks for the digital position. It remains unclear, moreover, how the abstract state differences that are assumed to underly everything (a view that is really a kind of *pythogoreanism*, much like the ancient view that "everything is number") are supposed to give rise to the physical structure of the world [129].

The project of digital physics would fit well with Solomonoff's theory, because it tells us that the bit sequences that SP relies on truly make up physical reality. The bit sequences that are the core of the model of SP then at once form an as-complete-as-possible representation. We could, nevertheless, hold fast to the conviction that the full information about the relevant events in the world can be exhaustively captured in a binary string ("reality can be fully represented by binary information"), while staying clear of ontological claims of any kind ("reality *is* binary information"). We would have to present grounds why this completeness of bit representation is reasonable, though, and in the process make much more precise what suppositions are needed to make such a picture work in the first place. One supposition that we would certainly be still bound to, and a very strong one at that, is a fundamental discreteness of nature.

Epistemic flexibility There is no reason, however, to enforce such a demanding requirement of completeness of bit representations. We are concerned with problems of prediction, and prediction has a distinctive *epistemic* ring. We *observe* a number of events in the world, or make a number of *measurements* of physical quantities, and we predict what we think we'll *observe* or *measure* next. I have purposefully put the characterization of predictive problems at the start of this Section in terms of *data*. This data is the result of direct observation or measurement of real-world events or quantities. It's this data that in turn has to be represented as a binary sequence.

Figure 4.1: From real-world events to observation data to a binary sequence.

The conceptual distinction between real-world events, observational data, and binary sequences is depicted in Figure 4.1.

Of course, any data we retrieve from the world must be finite, must have a maximum level of accuracy – which obviously implies discreteness at that level of precision. Every predictive problem could be naturally associated with some maximal level of accuracy of data. The final step from these discrete data items to a concatenated representation in a single binary sequence involves the choice of an *encoding*, a mapping from data chunks to binary codes, which introduces more freedom still. It's only sensible to allow for encodings at any desired level of fine-grainedness, at any level of abstraction. (Of course, for some predictive problems the data is already in binary form, and so data and representation simply coincide.) The supposition that it's always possible, in principle, to construct a single binary sequence via a given encoding looks perfectly reasonable. This gives a flexibility that fits the aim of a maximally unrestrictive model.

4.1.2 The Language of Binary Sequences

In getting rid of the demand or desire for a unique complete representation, and allowing for encodings at arbitrary levels of abstraction, we do find ourselves with a lot of freedom in choosing particular encodings. In fact, what constraints ought to be imposed – if any? Lacking constraints, we could adopt *any* encoding we like. This apparent absolute freedom in ways of coding is commonly held against algorithmic information theory, as indicating yet again that the endeavour must be highly *subjective*. If we could really pick any binary string of our liking to represent a particular observation or measurement value, how can claims about the algorithmic probability of these codes have any objective bearing on the data that we have, let alone the real-world phenomena the data are extracted from?

The issue could alternatively be phrased as the arbitrariness of picking a specific formal *language* to represent things. This brings to the fore the similarity of the current problem to the bane of Carnap's system – the dependence on a rather circumstantial formal language. One difference is that in our case, the choice of a language is much more unrestrictive. Another is that this choice doesn't really matter.

Information-theoretic grue To illustrate, consider a specific example by Kelly [95] to highlight the problem. He mentions an encoding that is inspired by Goodman's *new riddle of induction* [64]: a "grue-like" encoding that reverses assignment of bits to certain observations at specific times (e.g., assigning 0 to "green" and 1 to "blue" before time t, and the other way around from t on). The implication is that such arbitrary encodings must have an effect on the theory.

In replying to this example, let me be granted two things. First, there exists some "baseline" encoding such that there is an actual measure μ that assigns the "correct" probabilities to the data *via this encoding*. Second, the "arbitrary" reversals in the grue-like encoding follow an effective pattern (as in the onetime reversal at t).

Granted these things, my reply is that there must also be a measure that gives the same predictions based on this grue-like encoding as the assumed actual measure that presupposes the baseline encoding without reversals. Consequently, even on adopting such an outlandish encoding, the actual measure is taken into account by the universal predictor. There will certainly be variations within constant multiplicative bounds of predicted values, but to worry about this would be to fall in the same trap as wondering which particular universal predictor to use: the relevant thing is the existence of a class of universal predictors, and each member considers all such encodings.

Encodings and semimeasures Thus the threat of arbitrariness by encodings is taken away in much the same way as the threat of subjectivity by having to select a particular universal predictor – indeed, the two threats can be seen as two manifestions of the same issue of subjectivity. They are taken away by keeping to the level of the class of universal predictors. We may conclude that Solomonoff's language of bit strings escapes the most important objection against Carnap's formal language.

Admittedly, the above argument for the irrelevance of the particular encoding only works under two conditions. In the first place, it only applies to encodings that are effective variations of the baseline encoding. The theory won't be able to cope with an encoding that employs infinitely many reversals at completely random times. But it seems only fair to impose some minimal requirement of consistency of encoding, and effectiveness is the natural constraint in the current setting.

In addition, talk about variations presupposes the existence of a baseline encoding, an encoding that fits the actual measure at work. This supposition derives from the theory's starting assumption that *there is an actual computable measure μ that assigns the correct probabilities to the data*. Actually, a measure can only assign probabilities to binary strings (codes), so the assumption of an actual measure already includes a preferred "actual" encoding that acts as an intermediate between the probabilities and the data. See Figure 4.2, depicting that a measure indirectly assigns probabilities to the data via the direct assignment of probabilities to binary strings. Another way of looking at things, that is consistent with this general picture, is that there is a *class* of different actual measures, each with an associated encoding that makes sure the assigned probability values wind up at the right data.

In any case, the encodings are strongly related to the measures that are supposed to govern the data. The measures, as models of predictive *environments*, are the topic of the next Section.

4.2 Environments

The suggestive term *environment* serves to indicate that the probability measure gives a complete specification of the (probabilistic) constraints on the development of the

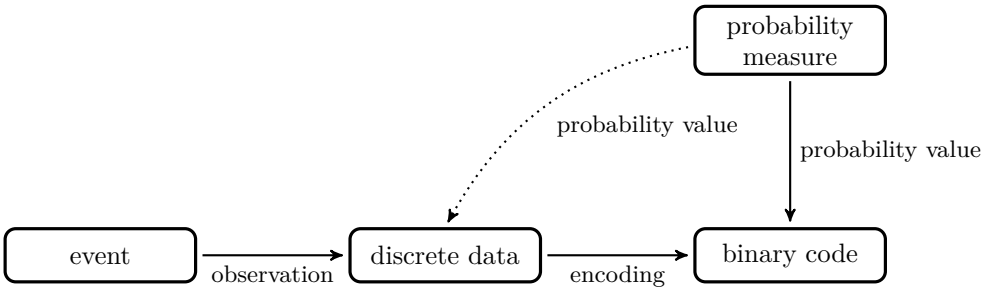

Figure 4.2: The assignment of probability values to data via codes.

(indeterministic) history. What concerns us in this Section, is how general this choice of probability measures is: could every real-world environment (that is, a specification of the behavior of a developing sequence of data) be modeled as a probability measure?

4.2.1 Generality of the Probabilistic Environments

It's hard to see how to come up with a supposition less restrictive than that of some probabilistic source at work – at least, if we are prepared to make the minimal assumption that the generation of a data sequence is governed or described by *something*. Effectiveness aside, there is no restriction whatsoever imposed on the probability measures, about independence or otherwise. Needless to say, this class includes *deterministic* sequence generation, that arises from a distribution that only assigns probabilities 1. Finally, recall that the measures have as sample space the class of finite strings, rather than $\{0, 1\}$, which means they provide for "time-dependent" probability values. It would seem that if there is *some* way of mathematically describing the generation of a bit sequence, then it must also be described by one such probability measure.

4.2.1.1 Measures or Semimeasures

The class $\mathcal{M}$ that is the basis of Solomonoff's theory actually involves *semi*measures rather than measures. As noted at the end of Section 1.3 of Chapter 1, however, the choice to let the family of all possible actual distributions extend to the class of semi-computable semimeasures would in the first place be motivated by the mathematical consideration that to have the class contain its own universal elements is the most elegant choice. This doesn't appear to be sufficient ground to bar us from simply restricting the possible actual distributions to the sufficiently general class $\mathcal{M}^{\text{meas}}$ of proper measures. The restriction of the proof of the convergence of universal predictors (Theorem 1.9, page 39) to actual measures gives us one reason to do so. Another reason is, of course, the rather odd notion of a "defective distribution", that yields probabilities that don't quite sum to 1, or that assigns a probability to a disjunction of events that surpasses the sum of the probabilities it assigns to the separate events.

The only consequence of keeping to proper measures is that the universal predictor

is not in that class, and might be said to be a bit *too* powerful: it can be seen to take into account distributions (namely, the strict semimeasures) that we excluded beforehand ourselves. Furthermore, one might feel uneasy about the fact that this predictor is a strict semimeasure itself. While we can plausibly refuse to admit strict semimeasures as actual distributions, we are inevitably left with universal predictors that are "defect". Should we worry about this?

Normalizations Solomonoff himself wasn't happy with the idea of a defective algorithmic probability distribution. He insisted on the "real" probabilities involved in the experiment with the universal monotone machine on random input (of the definition of algorithmic probability). What we are interested in, he maintains, is the actual relative probability that the universal machine in this setting produces a bit 0 rather than 1 after we have seen it produce a string σ. This comes down to the ratio $Q(\sigma 0)/Q(\sigma 1)$. As I discussed in Chapter 1, Section 1.3.2, the machine may produce *no* input[1] after σ: this causes $Q(\sigma) > Q(\sigma 0) + Q(\sigma 1)$, and to keep to Solomonoff's goal, the additional probability involved should be distributed over the possibilities of 0 and 1 in such a way that the ratio remains the same. To that end Solomonoff devised a unique *normalization* that turns the semimeasure into a proper measure [165]:

$$\begin{aligned} Q_{\text{norm}}(\epsilon) &:= 1, \\ Q_{\text{norm}}(\sigma 0) &:= \frac{Q(\sigma 0)}{Q(\sigma 0) + Q(\sigma 1)}. \end{aligned}$$

This *normalized algorithmic probability* indeed satisfies $Q_{\text{norm}}(\sigma 0)/Q_{\text{norm}}(\sigma 1) = Q(\sigma 0)/Q(\sigma 1)$ for all σ. However, the probabilities given by Q_{norm} are not even computably approximable anymore: Q_{norm} is not semicomputable. [119, s. 4.5.3]

Accepting semimeasures Even if we would go along with Solomonoff attributing central importance to the experiment of the universal machine, it's not clear at all why this would make us desire a ratio-preserving normalization that gets rid of the obsolete "non-halting" probability. Nor does the argument that the ratios of a proper measure really "*need* to be used in applications" [119, p. 304] appear to have much basis – certainly not if we don't go along with the method interpretation. That leaves us stuck at the starting point: that we simply have to preserve the "sacrosanct" notion of measure.

In lack of independent reasons to do so, I don't see why we should. The universal predictor's regularly assigning some probability measure to "no future" sounds undesirable, if put in this slightly provoking way, but at least there are ways of making sense of a prediction method that (mal)functions in this regard (something which would be more doubtful in the case of environments). It doesn't diminish in any way the predictor's single important property of formal universality, and so of its convergence properties. Indeed, it's formally *necessary* for universality – at least, if we want to uphold the requirement of semicomputability. For the only way to proceed otherwise is to introduce the normalized measure Q_{norm} that is no longer semicomputable. And this loss of semicomputability seems to me a far greater drawback than the use of a

[1] Solomonoff devoted one of his last papers [171] to the derivation of an upper bound on the probability of this occurring.

semimeasure. Effectiveness is a natural property of a predictor, and so the minimal level of effectiveness guaranteed by semicomputability, or effective approximability, *is* a desirable property of the universal predictor (also see Subsection **??**). The theory's result that *in a universal setting, there are universal predictors* is only stronger the more powerful the properties of the universal predictors, favouring a low threshold of effectiveness.

4.2.2 Interpretation of Probabilities

The probabilities of Solomonoff's framework are logical in the sense that they derive from purely formal stipulations. Similar to Carnap's approach, their values are simply defined by the choice of a probability measure; similar to Carnap's approach, the degree of confirmation between evidence and hypothesis (in SP, between a past and a future), mediated by Bayes' rule, follows from this choice. And similar to Carnap's approach, this leaves open the question what the probabilities should *mean*.

4.2.2.1 The Actual Measure

In regard to the question of the interpretation of probabilities, it makes sense to distinguish the two distributions that play a part in the setting of SP. There is the actual measure, and there is the universal predictor, a semimeasure.

The actual measure, as part of the universal model that is to accommodate any predictive problem, is to represent any actual environment. The chances involved in the probabilistic data generation are then given by this measure. What *kind* of probabilities are involved must fully depend on the process to be modeled, on the specific predictive problem. The universal model credibly suits any predictive problem, at any level of abstraction, and with any associated interpretation of probability values. Each predictive problem is bound to bring with it a natural interpretation of the probabilities involved: the flexible model of SP itself leaves this open.

4.2.2.2 The Universal Predictor

For every specific problem, the universal predictor provides a close approximation to the unknown actual measure. It simply approximates the "true" probability values, no matter how they are interpreted. Again, the matter of the meaning of the probabilities is deferred to the specific predictive problem accommodated.

However, there is another way of looking at the universal predictor, that is closer in spirit to Carnap's endeavour. While this presents a departure from the interpretation of SP that I pursue in this thesis, it connects to the main theme of universality.

Objective-logical restrictions I presented Solomonoff's theory in Chapter 1, Subsection 1.1.2 as a successor of Carnap's project of establishing objective probability values from logical constraints. The apparent tension between the objective-logical approach and the subjective interpretation of probability that Carnap inclined to is resolved by taking the former as an ideally rational limit case of the latter. In Carnap's own words,

> "I think there need not be a controversy between the objectivist point of view and the subjectivist or personalist point of view. (...) Basically, there is merely a difference in attitude or emphasis between the subjectivist tendency to emphasize the existing freedom of choice, and the objectivist tendency to stress the existence of limitations." [90, p. 119]

As noted by Zabell [189], the crucial point is the status of these limitations. Carnap believed that it should be possible to design a formal system that can impose restrictions that have objective and, indeed, *universal*, status, and that would finally result in unique probability values.

The position of the opposing subjectivist camp is that such restrictions can never be more than auxiliary stipulations. These might very well be employed in certain circumstances but won't hold in others. Universally valid and logically compulsory constraints, of the kind envisaged by Carnap, do not and cannot exist.

The universal prior distribution Recall from Chapter 1, Subsubsection 1.1.2.2 the Bayesian problem of agreeing on the prior probabilities. I originally introduced universal $\mathbf{M}$ not as a prediction method, but as a universal prior distribution. In line with Carnap's directions, this prior distribution then also arises from purely formal constraints. It can be seen to result from an "objectivist tendency to stress the existence of limitations". These limitations are the limitations that are under discussion in this Chapter – the constraints of a language of bit strings, and of probabilistic and effective environments.

At the same time, the purpose of this Chapter is to show how unrestrictive these constraints are. Sections 4.3 and 4.4 will impress more clearly still how the constraint of effectiveness is sufficiently strong to allow us to come to nontrivial results, while still being sufficiently mild to hardly put us in danger of excluding anything of interest. In the case at hand, the restrictions are sufficiently strong to allow us to pin down a prior distribution, while sufficiently mild to still pertain to full generality.

To skeptics of the possibility of universally objective constraints, the main proof of the failure of Carnap consists in the arbitrariness of the artificial language that he had to employ. I argued in Subsection 4.1.2 that Solomonoff's language of binary strings evades this objection. However, notwithstanding the pronounced generality of the language of bit strings, the final dismissal of the problem of subjectivity relied on my interpretation of the theory of SP as asserting *the existence of a class of universal predictors* in a universal setting. Elevating universal $\mathbf{M}$ to the status of a *unique* universal prior distribution, we would again have to face the issue of subjectivity (Chapter 2, Section 3.2) in earnest: there are in fact infinitely many asymptotically equivalent universal prior distributions, none preferable to the others. The moral seems to be that the objective-logical approach based on Solomonoff goes a long way towards the definition of a universally valid prior distribution, but not all the way: much to the liking of the subjectivist, there remains a strip of freedom that continues to resist invasion.

4.3 Effectiveness

The framework of measures on encodings of bit strings leaves little to wish for in terms of generality. Of course, that's not the whole story: the perfectly general measures are still constrained by the requirement of *effectiveness*.

4.3.1 A Natural Restriction?

The choice for an effective setting is obviously a restriction to the model's generality. Unlike the bare probabilistic setting, it's clear that the effective setting leaves things out: distributions that are not computable. However, in line with the way I presented the choice for this restriction in Chapter 1, Section 1.1.3, we must evoke some restriction to find grip, to distill anything nontrivial. It simply seems impossible to handle full-blown generality – where to begin?

Algorithmic randomness: effectiveness to the rescue A nice analogy to the justification of a restriction to effectiveness is found in the inception of the field of algorithmic randomness (that we encountered before on page 74). The field originated in the goal to provide a definition of *randomness* of infinite binary sequences. There are several intuitions connected to the notion of randomness, that can all be associated informally with binary sequences. A random sequence should *lack regularity*, making it hard to compress. It should be *unpredictable*, making it hard to make money by placing bets on its continuation. But it should also be *typical* in the sense that it shows no statistical anomalies. The problem is how to make these intuitions precise. Indeed, how to capture something that is characterized by evading characterization in a precise characterization? A sequence can't have no structure *at all* – not if randomness is itself a structural property. And a sequence can't show no statistical atypicalities *at all* – not if the intersection of all classes of typical sequences turns out to be empty. Full generality, in this case, is simply contradictory.

The solution is again the restriction to effectiveness. A random sequence is stipulated to be a sequence that is *effectively* incompressible (all initial segments have maximal Kolmogorov complexity). There are no *effective* betting strategies that allow you to make money on its continuation (there is no semicomputable *martingale* that yieds unbounded capital along the sequence). And it passes each effective test for atypicality (it passes each *Martin-Löf test* [121]). Further credibility to this move is given by the result that those three requirements isolate the exact same class of random sequences – though we have seen that there is still some amount of freedom in choosing the proper level of effectiveness (page 74).

If some restriction is unavoidable, the restriction to effectiveness looks very mild indeed. All that's imposed, on the face of it, is that things are calculable in principle – supposing that this is accurately captured by Turing-computability. As a matter of fact, the full generality of the effective setting is oftentimes simply defended (if at all) by a direct appeal to the Church-Turing Thesis:

> "It should be appreciated that according to the Church-Turing thesis, the class of all computable measures includes essentially any conceivable natural environment." [140, p. 1118]

Unfortunately, things are a bit more complicated.

4.3.2 Turing Computability in the Wider World

Let's focus for the moment on the general question whether real-world processes are adequately modeled by Turing-computable functions. The claim that they are, can actually be broken down into two separate claims:

1. Real-world processes can be considered computable in the intuitive sense (calculable).
2. Informal computability (calculability) is adequately characterized by Turing-computability in the formal sense.

The Church-Turing Thesis (CTT) underwrites the second claim – provided that there is agreement on the original sense of "calculable". This is not the case, resulting in a distinction between what we may dub the *Original* CTT and the *Physical* CTT. The *Bold* version of the latter also extends to include the first claim.

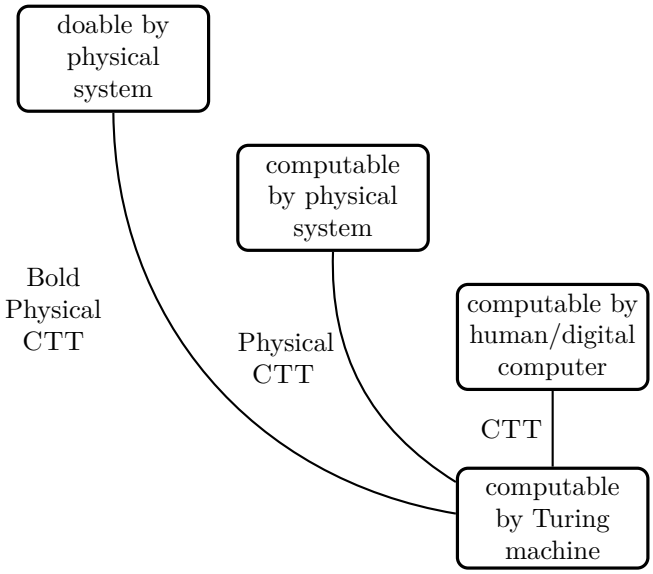

Figure 4.3: The various versions of the Church-Turing Thesis.

4.3.2.1 The Original Church-Turing Thesis

Recall from Chapter 1, Section 1.1.3 that the Church-Turing Thesis (CTT) states that the informal notion of "effective method" should be identified with the formal notion of "(Turing) computability". As B.J. Copeland [31, 29] emphatically argues, the notion

of effective method originally pertained to the type of procedure that an idealised human computer with only pen and paper and unlimited patience could perform. In the words of Wittgenstein:

> "Turings »Maschinen«. Diese Maschinen sind ja die *Menschen*, welche kalkulieren." [186, §1096]

The basic steps such a dulled human computer is allowed to perform – and the same holds for the basic operations a digital computer is able to perform – can clearly be simulated by a Turing-machine (see also Chapter 1, Section 1.1.3), and so the content of this interpretation of the CTT is rather uncontroversial. But it's also a bit limited: too limited in any case to function in the way it is regularly employed. Very often (the quotation on the previous page being just one example) the CTT is called on to justify the convenient move of treating *any computation at all* at all as Turing-computable. If we wish to preserve the original meaning of "effective method", however, such a move requires something stronger than the original CTT.

4.3.2.2 The Physical Church-Turing Thesis

When we talk about being computable *at all*, in our world, we are talking about the extent of *physical* computability. The associated version of the CTT, that *any physically computable function is Turing-computable*, can be branded the *Physical Church-Turing Thesis* (following [131]). An equivalent formulation is given by R.O. Gandy's [58] *Thesis M: whatever can be calculated by a physical machine is Turing-computable*.

The Physical CTT is stronger than the original, because there may exist operations that can be done by a physically possible machine but that transcend effectiveness in the original sense. If Turing-computable functions are indeed limited to the effectively calculable functions, the existence of such Turing-incomputable operations, while leaving the original CTT untouched, would refute the Physical CTT.

Hypercomputation Computation of Turing-incomputable functions is nowadays called *hypercomputation* (following [30]). While there are many models of hypercomputation, they are in the first place purely theoretical, *notational*, models, and it remains to be seen if any of them is also physically realizable. Examples of such hypercomputation models are infinitely accelerating machines (each operation is performed twice as fast as the previous [28]) and infinitely shrinking machines (each operation is followed by the machine's producing a self-copy that aids in the computation [35]); models that are more than likely physically impossible. Sometimes *quantum computers* are naively hailed as going beyond Turing-computability, but while such computers can be more efficient, they can't compute more things.

A related line of attack is the investigation of the possiblity of sources of incomputability in the physical world, that could potentially be harnessed as operations of a hypercomputation machine. Turing's [173] original model of *oracle* computation, that allows for consultation of an incomputable real, has recently been used (misused, to some[2]) to reinforce speculation

[2] Copeland's presentation of his project of hypercomputation as a legacy of Turing himself, particularly his and D. Proudfoot's exposition in *Scientific American*, has attracted heated response from other Turing scholars. The controversy even found its way within the *Stanford Encyclopedia of Philosophy* when A. Hodges circulated a letter criticizing Copeland's article as well as his En-

[30] that the physical universe may contain incomputable information that could feasibly be consulted as an oracle by a hypercomputer. Often mentioned in this context are the specifications of M.B. Pour-El and I. Richards [132, 133] of classical physical systems that are given by equations with incomputable solutions – although it's still unclear whether such situations can occur in reality. Also noteworthy is H.T. Siegelmann's presentation in *Science* ([158], see also [159]) of *neural networks* that could supposedly compute more than Turing machines. What these and related speculations have in common is their reliance on the exploitation of incomputable real numbers, and even if current physics tells us that many physical quantities can take arbitrary real values and we know for a fact that almost all reals are incomputable, there is no evidence at all that it's possible to measure such reals with unbounded precision, and so fully utilize them in computation [36, 37].

Computations and processes The Physical CTT concerns the extent to which Turing machines capture the intuitive concept of a physical "computation". But what physical processes do we count as computations to begin with? In a similar vein, what physical systems can count as machines?

A sensible approach is to dissect our intuition into a number of requirements like repeatability, resettability and readable input and outputs (cf. the *mechanistic* account in [129, 130]). Although it has the shortcoming of remaining a bit imprecise, this approach connects well to a practical understanding of computation. For that reason, in spite of Copeland's belief that we should be "profoundly surprised if the physics of the real world can be properly and fully set out without departing from the set of Turing-machine-computable functions" [31, p. 64], the conclusion of the overview in [129] is that "for the time being, (...) Physical CTT remains plausible. It may well be that *for all practical purposes*, any function that is physically computable is Turing-computable [emphasis mine]."

The other extreme is the pancomputationalist view that *everything* is in fact a computation, simply because any physical system can be mapped onto a computational description (as promoted by Putnam [136]) or because the universe is itself a computer (a view that we have encountered before on page 84). The proper conception of "computation" remains a complicated issue, and we can try to circumvent it altogether by skipping the intermediate step of "effectiveness" or informal "computability", and undoing the distinction of two separate subclaims that I started the current discussion with. In other words, we can directly consider the claim that *any physical process is Turing-computable*. This stronger claim that anything *doable* at all by a physical system is computable in the formal sense is called the *Bold* version of the Physical Church-Turing Thesis in [129].

The Bold Physical Church-Turing Thesis Unfortunately, this general statement of the Bold Physical Church-Turing Thesis can also be made precise in a number of very different ways – an echo of the difficulty of making precise the notion of computation, that we just hoped to have left behind.

One obvious interpretation, that stays close to the idea of the Turing machine as representation of a *function*, an *extensional* input-output relation, requires that every observable

cyclopedia lemma [29] about the CTT, after which Hodges was invited to write his own lemma about Turing [71], subtly counterbalancing Copeland's interpretation.

value of a physical system is Turing-computable as function of the time [131]. Another, that is probably closer to the idea of "doable", demands that every physical process can be *simulated* by a Turing machine, much like the steps of a Turing machine were originally intended to simulate every step of a human computer. This interpretation has a close kinship to the *Church-Turing-Deutsch principle* [44] that a universal computer can simulate any physical process, and also to what Copeland [27] calls *Thesis S*, that any process that is mathematically describable or scientifically explicable can be simulated by a Turing machine (regularly instantiated in the philosophy of mind as the supposition that the human brain may be modelled by a Turing machine; e.g. [154]).

Regardless of the exact interpretation of the Bold Physical CTT, our extending the scope of the Physical CTT from informal "computations" to "processes" exposes it to a straightforward refutation. What we lose in this step is a restriction of finiteness (regarding objects that can be read or manipulated) that is only natural for the concept of computation, but not necessary so for the concept of physical process. This removes the barrier against the hope of most hypercomputation models: the presence or manipulation of real-valued quantities in physical processes. As most reals are incomputable, it would be impossible to maintain that a Turing machine can simulate such processes or calculate the values of these quantities.

So we see that this matter for large part reduces to the very much open question of the fundamental discreteness of physical reality. To those who would answer that question in the negative, the Bold Physical CTT is fairly trivially refuted.

4.3.3 Effectiveness in Solomonoff Prediction

Now that we have a somewhat clearer view on the actual range of the various versions of the CTT, I proceed to the issue that really concerns us: what claim properly expresses the validity of the effective model of SP?

4.3.3.1 Randomized Monotone Machines

Finding inspiration in the previous theses, we should like to express a conjecture about the link between the effective operations of our model and effective operations in the real world. The conjecture we look for concerns the relation between real-world environments and the effective environments of SP: the computable probability measures.

Actually, for the benefit of discussion, it will be easier to consider the more general case where we don't restrict the class of environments to the proper measures in $\mathcal{M}^{\text{meas}}$. Instead, we consider the weaker statement about lower semicomputable semimeasures in $\mathcal{M}$ – the important thing is that this statement is still a necessary part of the stronger version involving only computable measures.

Specifically, we desire that every real-world environment corresponds to a semicomputable semimeasure. We desire that the probability measure over data sequences induced by any physical process corresponds exactly to some semicomputable semimeasure over these data sequences.

However, the effectiveness of these measures pertains to the potential calculation of probability values, and it's not entirely obvious how we should like to map these potential calculations to processes in reality.

Luckily, we have an exact correspondence between lower semicomputable semimeasures and monotone Turing machines working on uniformly distributed input (Chapter 1, Subsubsection 1.3.2.1). This provides us with a much more concrete picture. Every environment should then be given by a monotone machine that receives uniformly distributed random input. More figuratively yet: a monotone machine that has a (uniform) random number generator at its disposal. The latter move has the added benefit of doing away with the demand of a steady supply of input, and the question where this input would come from.

Thus we desire that every history is generated by some *randomized* monotone machine.

4.3.3.2 A Tailor-Made Church-Turing Thesis

The matter at hand, then, is whether every real-world history of observed data can be seen to be generated by a process in a way that is effective as formalized by randomized monotone Turing machines. This is conveyed by the following statement – let's call it the *Church-Turing-Solomonoff Thesis* (CTST).

Thesis 4.1 (Church-Turing-Solomonoff Thesis)**.** *The data generation capabilities of any physical system are given by the data generation capabilities of some randomized monotone Turing machine.*

Here I employ the expression "data generation capabilities" to account for the possibly indeterministic character of the data generation process: it refers to the probabilities for each data sequence of being generated. The thesis states that the probabilities on data sequences induced by any physical system correspond to the probabilities induced by a randomized monotone machine.

The CTST expresses what we have to assume about the scope of Turing computability with respect to the model of SP, if we hold that the restriction to effectiveness is in fact not a restriction at all. Let us now turn to the question how credible it is.

A first evaluation Note first that our CTST has in common with the Bold Physical CTT that it doesn't bother with an intermediate informal notion of "computation": it's a single bridging claim about the relation of the physical world with SP's formal apparatus. This would also seem to expose the CTST to the objection regarding real-valued quantities in the physical world. A physical process that generates real-valued values for physical quantities must leave any monotone machine behind.

This vulnerability to a continuous physical reality should be familiar by now – as should the way to address it. It's counteracted by the consideration of *data* rather than real-world events or quantities. When we consider the data generated (or generable) by a physical system, this includes our imposing *some* finite maximal level of accuracy: this discrete level of coarse-grainedness also applies to the data (including the

additional but inconsequential encoding into bit strings) generable by any monotone machine we might want to correspond to the system.

For this reason, our CTST is much less demanding than the Bold Physical CTT. Note that the CTST also only asks for an *extensional* correspondence between generable outputs, which makes it weaker than the interpretation of the latter that asks for an actual *simulation* of the physical process by the computation steps of the formal machine (with all the additional unclarities concerning the proper mapping).

Another respect in which the CTST is better equipped is the allowance of a random source. This makes the monotone machines of our thesis more powerful than the regular Turing machines that the Bold Physical CTT covers. See, for example, the discussion in [7] of a computer that employs a source of quantum randomness to generate Turing-incomputable sequences. Such a computer might be a counterexample to the Bold Physical CTT; not so for the CTST.

Back again However, the allowance of random sources in physical systems opens up another origin of continuity. Namely, such a source may embody an incomputable distribution, and so induce incomputable probability values on the system's data sequences. Our randomized monotone machines, on the other hand, can only manipulate the values from a uniform source, resulting in probability values on data sequences that are at least (semi)computable. This origin of incomputable continuous values is not so easily dismissed.

It appears we're thrown back at our main question in its original, undiluted form: how much of a restriction is the requirement of semicomputable probability values, that is, how much of a restriction is the semicomputability of the semimeasures? How much of a restriction is the computability of the measures?

The sobering result of our detour via various versions of the CTT is a further proof of the circumstance that the potential restrictiveness of the effective model of SP boils down to one aspect about which little more can be said: the possibility of physical sources of incomputable probabilities. The plausibility of the normal Physical CTT, disqualifying Turing-incomputable physical computations, transfers to the plausibility of the generality of the monotone machines as physical data-generating computations; the avoidance of the main threat of continuity to the Bold Physical CTT speaks well for the generality of the monotone machines as physical data-generating systems. What remains is the open empirical question of the existence of sources of incomputable probability values in nature.[3]

The conclusion must be that what we face here is a genuine threat to the universality of the model. In the next Section, I propose a way of answering this threat.

[3] Note that the generation of a sequence with the help of an incomputable random source, as a threat to the normal Physical CTT, can be dismissed on the grounds that the Physical CTT (unlike the Bold Physical CTT) rests on the informal notion of "computation", and it's questionable that this notion includes random generation of a sequence.

4.4 Predictors For Environments

Up to this point, I have interpreted the probability measures in $\mathcal{M}^{\text{meas}}$ as *environments* that govern the data generation. But as mentioned in Chapter 1, Subsection 1.4.1, equally valid is the interpretation as *predictors*.

4.4.1 The Model of Predictors

This alternative interpretation realizes a retreat to the purely epistemic domain, the domain of our methods of prediction. Importantly, we do away with the actual "true" environment μ that was supposed to govern the generation of the data.

As an added benefit, the new interpretation naturally extends to the complete class $\mathcal{M}$ of semicomputable semimeasures.

4.4.1.1 Predictors as Semimeasures

In most of the previous, the range of possible environments was restricted to the subclass $\mathcal{M}^{\text{meas}} \subset \mathcal{M}$ of computable measures. One reason for this choice is that the convergence Theorem 1.9 only holds for environments that are proper measures. (Recall from Chapter 1, Subsubsection 1.3.1.3 that the notions of semicomputability and computability coincide for measures.) Another is that the notion of a real-world environment as a semimeasure is hard to make sense of.

As hinted at in Subsubsection 4.2.1.1, the notion of a prediction method as a semimeasure is not quite so mystifying. While it's unclear what to make of a process that is governed by a defective distribution, there is no reason why we couldn't construct a method of prediction that returns probabilities that don't exactly sum to 1. Indeed, the universal predictor *must* be a strict semimeasure. That means that we now do have good reason to let the class contains its own universal elements: they are all predictors. It is a reason to take the complete class $\mathcal{M}$ as our class of possible predictors.

This is a nice consequence of the new interpretation. Whereas in the previous interpretation the universal predictor took into account elements that in fact weren't environments at all (namely, the strictly semicomputable semimeasures), the dominance of the universal predictor now concerns precisely every possible predictor.

4.4.1.2 Dropping the Actual Environment

If all that's left to us are prediction methods, we are forced to remain silent on the possible origins of the data. It's fine to suppose that there exists a most *informed predictor* ρ (or a family of such predictors), that, as a semimeasure, is closest to the actual source of the data, and so can be expected to give the most accurate predictions. In a way, this informed predictor takes the place of the actual environment. However, the two shouldn't be identified: the informed predictor is a semicomputable semimeasure, but we no longer assume that the actual source of the data can also be described in this (effective) way.

Completeness of the universal predictor The formal universality of the universal predictor consists in the fact that it formally dominates every member of $\mathcal{M}$. In the current interpretation that means that it formally dominates every other predictor. Informally, it takes into account every possible predictor. The completeness results of Chapter 1, Subsection 1.4.2 no longer concern convergence to the "truth" of actual μ: they state that the universal predictor quickly converges to the predictions of the (computable) informed predictor ρ.

In fact, we can establish a more refined result. We can derive a bound on the difference in predictive success of the universal predictor $\mathbf{M}$ and the informed predictor ρ on an unknown environment μ. We don't need to assume anything about the effectiveness of this measure μ: it might not even be semicomputable. Moreover, we don't need to assume that ρ is a proper measure: it can be any semicomputable semimeasure.

The measure of predictive success or expected error that I use is the summed Kullback-Leibler divergence of the predictor with respect to the environment.[4]

Theorem 4.2 (Refined convergence in difference)**.** *For arbitrary actual measure μ and semicomputable semimeasure ρ, the additional expected total prediction error when predicting with $\mathbf{M} = \xi_w$ instead of ρ is*

$$\sum_{t=1}^{s} D_t(\mu \parallel \mathbf{M}) - \sum_{t=1}^{s} D_t(\mu \parallel \rho) \leq -\ln w(\rho).$$

Proof. Mimicking the steps (a) to (e) in the proof of Theorem 1.9, page 39, we derive that the above left-hand side equals

$$\sum_{|\sigma|=s} \mu(\sigma) \ln \prod_{t=1}^{s} \frac{\mu(\sigma_t \mid \sigma_{<t})}{\mathbf{M}(\sigma_t \mid \sigma_{<t})} - \sum_{|\sigma|=s} \mu(\sigma) \ln \prod_{t=1}^{s} \frac{\mu(\sigma_t \mid \sigma_{<t})}{\rho(\sigma_t \mid \sigma_{<t})}.$$

(Note that in step (d) we require that μ is a proper measure, but no such assumption is needed for $\mathbf{M}$ or ρ.)

This term then reduces to

$$\sum_{|\sigma|=s} \mu(\sigma) \left(\ln \prod_{t=1}^{s} \frac{\mu(\sigma_t \mid \sigma_{<t})}{\mathbf{M}(\sigma_t \mid \sigma_{<t})} - \ln \prod_{t=1}^{s} \frac{\mu(\sigma_t \mid \sigma_{<t})}{\rho(\sigma_t \mid \sigma_{<t})} \right) = \sum_{|\sigma|=s} \mu(\sigma) \ln \prod_{t=1}^{s} \frac{\rho(\sigma_t \mid \sigma_{<t})}{\mathbf{M}(\sigma_t \mid \sigma_{<t})}.$$

Unravelling steps (f) to (i), using the dominance of $\mathbf{M}$ over $\rho \in \mathcal{M}$ within constant $w(\rho)$, then leads to the bound $-\ln w(\rho)$. □

This shows that the universal predictor almost always quickly predicts as well as any competing predictor, no matter the actual (possibly not even semicomputable) environment. The fact that ρ can be any semicomputable semimeasure corroborates our extension of all possible predictors to the complete class $\mathcal{M}$.

[4] A similar result could be derived in terms of the squared Hellinger distance, but this requires a very elaborate proof [P. Grünwald, personal communication].

4.4.1.3 Solomonoff Prediction and the Problem of Induction

We can now address the question that is the title of [168]: *Does Algorithmic Probability Solve the Problem of Induction?* According to Li and Vitányi,

> "R.J. Solomonoff's inductive method (...) may give a rigorous and satisfactory solution to this old problem in philosophy." [119, p. 347]

The problem Hume's original challenge concerned the *justification* of induction. The argument[5] is that inductive inferences are characteristically *underdetermined* by the premises: there are different conclusions, different ways of amplifying our knowledge, that are just as well consistent with the premises. Hume suggests that the underlying principle we use to justify our inductive inferences is the *uniformity of nature*: the "course of nature continues always uniformly the same" [74, p. 89]. However, a defence of this principle, along the lines that it has always been this way, hence will be so in the future, is itself an inductive argument. In general, a justification of induction can't be deductive (because of underdetermination) nor inductive (because of circularity): therefore no justification can be given [120].

Note that this point is strictly limited to the impossibility of giving an argument that justifies induction in the face of someone who doesn't accept induction – a situation that is no different for the completely uncontroversial procedure of *deduction*! The impossibility of giving an argument that justifies induction doesn't imply that induction is never sound [34, ch. 4]. Further assumptions or the restriction to specialized kinds of induction might also open the way for a partial justification. Nor does Hume's conclusion bar a slightly more pragmatic vindication by deductively showing that a self-corrective method of repeated induction will bring us closer to the truth. Finally, even if one can't prove that induction is *reliable* (always leading to the truth), there are ways of showing how certain inductive methods are *optimal* (the closest we can get to the truth) in some sense. Formal learning theory, that, as we have seen (page 65), classifies inductive methods according to their truth-convergence, exemplifies such a project.

The answer Solomonoff Prediction is not a solution of Hume's problem of induction, that, anyway, "in the absence of further assumptions (...) is and should be insoluble" [180]. Note that the problem of induction is essentially the same as what I called the problem of prediction before. The problem is that not assuming anything simply means that *anything can happen*, in which case every specific prediction has the potential to fail.

In the original interpretation of the model, we do assume something: that the actual environment is given by a measure that is effective. In directing SP against Hume, this starting assumption takes the role of the question-begging assumption of the uniformity of nature.

However, the purpose of the previous three Sections was to show that it is a very *mild* assumption. And it has allowed us to prove the existence of *reliable* methods of prediction: the universal predictors quickly converge to the truth. These two facts are conveyed by the statement that *in a universal setting, there are universal predictors.* In a setting that is universal in the sense that only very little is assumed, we can derive

[5] In modern terminology – Hume doesn't even use the word "induction" in his argument in the *Treatise of Human Nature.*

the existence of methods of prediction that are universal in the sense that they are (almost) always reliable.

If it were the case that the starting assumption is in fact not a restriction at all, then, since it still give us reliable methods of induction, one could perhaps claim that a solution to Hume's problem is in sight. But the current Section is motivated by the acknowledgement that the assumption of effectiveness might in fact be a substantive restriction on what can happen in the world.

In the new interpretation of the model, we do away with this assumption – and so with any reference to an underlying "truth". This also prevents us from saying anything about the reliability of predictors in our model, blocking a full-blown justification of their use or the use of induction in general.

What we now *can* say is that SP establishes the existence of predictors that are *optimal*. They signify the best we can ever do. In a setting that is universal in the sense that only very little is assumed, there are predictors that are universal in the sense that they are (almost) always optimal. As I will argue next, it is also plausible that in the new interpretation the starting assumption is *not a restriction at all*.

Theories of optimal prediction The existence of optimal predictors in a universal setting is also a point that may inspire more specialized theories of induction. One relevant approach that is inspired by Solomonoff is the theory of *individual sequence prediction* (or *prediction with expert advice*) [182, 14, 15]. Originally a theory of machine learning, it aims for prediction algorithms that minimize the worst possible prediction error by evaluating the performance of other predictors ("experts"). Characteristically, no stochastic assumptions regarding the data source are needed.

Much related is G. Schurz's project of *meta-induction* [152, 153]. A strategy of keeping an eye on the past track record of other prediction methods, and mimicking those that were most succesful, results in a prediction method that can be shown to be optimal in a universal sense. Again, there is little that can be said about the method's reliability (in some outlandish scenarios all predictors might be very bad), but it's guaranteed to be the best we can do. The position that optimality is a satisfactory justication of meta-induction can be seen as an echo of Reichenbach's [141, §91] best-alternative argument for induction. Indeed, it is argued by Schurz that this *a priori* justification of meta-induction leads to a non-circular justification of regular induction applied to objects or events.

4.4.2 Universality of the Model of Predictors

The main line of the current Chapter was to argue for the universality of the model of SP. Picking up this line again, our concern is the generality of the model when the measures are viewed as predictors instead of environments.

Replacing environments by predictors and so dropping the assumption of an actual measure doesn't affect the conclusions in Sections 4.1 and 4.2 on the generality of the information-theoretic language and the setting of probability measures. The data can be represented, at the desired level of abstraction, by a developing binary string; the predictors evaluate the possible future data by assigning probabilities to continuations of the string. The precise (effective) encoding is of no consequence since the universal

predictors account for every such encoding. And just like probability measures should be sufficiently general to represent environments, so should semimeasures certainly be sufficiently general to represent predictors.

What remains is the familiar constraint of effectiveness.

4.4.2.1 Effectiveness of Predictors

The discussion of the previous Section of the effectiveness of environments got stuck at the very much open question of the Turing-computability of probability values of random sources in nature. Transferring the restriction of effectiveness to the predictors, we find ourselves in a much better position.

Another tailor-made Church-Turing Thesis The claim about Turing-computability of processes in nature that was codified by the Church-Turing-Solomonoff Thesis turns into a more sober version about Turing-computability of our prediction methods. To be specific, the claim of interest is the following Modest version of the CTST.

Thesis 4.3 (Modest Church-Turing-Solomonoff Thesis)**.** *Every prediction method is given by a semicomputable semimeasure.*

It's now more natural to turn away from talk about data generation capabilities and directly address the semicomputability of the prediction methods as semimeasures, that is, the semicomputability of the probability values that our prediction methods yield.

Evaluation of the Modest Church-Turing-Solomonoff Thesis The claim that the operations of our methods of prediction must always be calculable almost looks self-evident. Contrary to the earlier thesis about processes in nature, we are considering the prediction methods we can design ourselves, and their description or implementation as useful methods basically involves as a minimal condition their being calculable. Indeed, we only require that the predictor's probability values are "semicalculable" they even allow for the mere possibility of calculation of increasingly accurate approximations from below. Of course, the further step is to identify this intuitive (semi)computability with Turing-(semi)computability.

However, strictly speaking, we would still need to cover in our thesis all computations that are at all physically possible. After all, we could, in principle, design prediction methods that harness such computations.

This observation can't take away the strong intuition that a thesis about our own methods is in fact a bit weaker than a thesis about physically possible computation. In order to give this intuition more basis, recall that the Physical CTT, that covers all physically possible calculation, still hinges on the scope of the intuitive concept of computation. We saw that in the extreme case, one can consider *every* process in nature a calculation. However, it seems much more fitting to try to demarcate an idea of computation that is closer to practice, a conception that involves requirements of repeatability and locality and the like. How to make this precise would take us

too far here; the relevant thing is that it would result in a most narrow notion of calculability. We saw earlier that, in lack of evidence for physical hypercomputation, the corresponding interpretation of the Physical CTT is downright plausible. In fact, as a limit of imposing reasonable restrictions on the notion of calculation, and also connecting to predictive practice, it's not unreasonable to confine ourselves to calculations that can be performed on a digital computer. In that case, all that we require is the uncontroversial original CTT. Then the constraint of computability, let alone semicomputability, is in fact no restriction at all.

The crucial point to note is that the previous argument would be highly suspect in the case of effectiveness of environments (as these involve real-world *processes* rather than our own calculations), while it is only natural in the case of predictors. Indeed, the assumption of effectiveness of predictors is in complete agreement with practice, while the assumption that environments are effective is in fact not: in statistical practice, more often than not the models under consideration have incomputable parameters [P. Grünwald, personal communication].

This renders the Modest CTST in the model of predictors much more convincing than the original CTST in the model of environments. The constraint of a minimal level of effectiveness – semicomputability – of predictors should hardly be a constraint at all. The epistemic setting of prediction methods is much more receptive to a restriction of effectiveness, the latter being itself very much an epistemic notion.

4.4.2.2 Conclusion

In addition to the universality of the language of bit strings and the perfectly general identification of predictors and semimeasures, we have just seen that effectiveness of predictors is also a completely reasonable and undemanding condition. We may conclude that the model can, in principle, capture all possible predictors in any given predictive problem.

One could argue that all possible competing predictors, exhibiting all that we can ever do, is also the very most that can still be of interest in any given predictive problem. We would be interested in the best possible, informed predictor, representing the best we can do; the further happenstance of an actual environment, that might not be effective and so escapes even the informed predictor, is strictly irrelevant because it is fundamentally out of our reach. This is a defence of the value of optimality rather than reliability.

The new interpretation of all elements of $\mathcal{M}$ as predictors has the benefit of being much more consistent than the former interpretation of the members of $\mathcal{M}$ as both environments (restricted to subclass $\mathcal{M}^{\text{meas}}$) and predictors (the universal elements in the subclass $\mathcal{U}_{\mathbf{M}}$). Most importantly, the new interpretation has the benefit of assuming much less. By retracting any claims on true environments we may replace the strong version of the CTT that environments must be effective by the much more credible supposition that our own prediction methods must be effective. If we accept the argument for optimality rather than reliability, then this means that the model is even more universal, to the point of being absolutely unrestrictive; in which case this is the best way of looking at the model of Solomonoff Prediction.

Conclusion

The essence of Solomonoff's theory of Prediction is that *in a universal setting, there are universal predictors.*

The mathematical framework Concepts from information theory, measure theory and computability theory inspire a general setting of prediction, as well as the definition of all-encompassing predictors within this setting. Formally, all predictive environments are represented by (strictly computable) elements of the class $\mathcal{M}$ of semicomputable semimeasures on bit strings, and this class contains elements that are formally universal in the sense that they dominate all others. Subclasses of such universal elements are isolated by the two seemingly different definitions of the algorithmic probability Q and the universal mixture distribution ξ. In fact, these two definitions appear to give the exact same class of asymptotically equivalent semimeasures. Moreover, one can prove that elements of this class quickly converge to any other computable measure. These facts are motivation to talk about a single universal predictor $\mathbf{M}$.

Interpretations The question is what to make of this strictly formal framework. This question about its interpretation relates to the question about the purpose of Solomonoff Prediction. I distinguish three broad ways of looking at SP and its aims: as a method, as a model, and as an actual theory of prediction.

Solomonoff's results are often heralded as handing us an idealized yet universally objective method of prediction. I reject this view, on the ground that the bare fact of an idealized method has neither practical nor theoretical use. The model interpretation, too, must give way in light of the extreme amount of idealization involved: the framework is simply too abstract to gain us any useful insights by modelling predictive problems.

This leaves the interpretation of SP as a theory, that I simply take as a (collection) of statement(s) that provide a broader context for a model and/or a method. The aim of such an idealized theory must either be providing a better understanding of related (philosophical) issues, or serving as inspiration for more practically oriented theories. Both these aims are supported by the identification of fundamental underlying ideas, of core principles.

Taking a lead from the heuristic ideas that are commonly associated with the theory, I presented three candidate core principles. Those principles are, first, Completeness, expressing that the convergence results regarding the universal predictor is sufficient to show that the theory works, second, Simplicity, highlighting the universal predictor's supposed bias towards simple hypotheses, and, finally, Universality, with the

connotation that everything is taken into consideration.

My defence of the primacy of the principle of Universality goes hand-in-hand with my proposed interpretation of the theory of SP, as boiling down to the statement that in a universal setting, there are universal predictors.

The problem of prediction The philosophical context for my interpretation of SP is the problem of prediction: in the absence of restricting assumptions, *anything can happen*. The more restrictions we enforce on the environment (at increasing risk of losing generality), the more we can conclude about the future. The best possible way out of the problem of prediction is to somehow introduce assumptions that are minimally restrictive, but that allow us to be maximally determinate about the future.

It's natural to rephrase the latter half as having a method of prediction that always (quickly) produces predictions that are maximally close to the truth. If the environment is indeterministic, then this involves the true probabilities over all possibilities.

The problem of prediction must be seen as underlying the statement that in a setting that is universal in the sense of minimally restrictive, there are predictors that are universal in the sense of almost always convergence to the true future.

In a universal setting... The model of SP adheres to Universality because it is primarily aimed at preserving maximal generality. Every real-world data-generating predictive environment can be expressed as a probability measure on finite bit strings. The model's information-theoretic language resists in my interpretation the standard "grue" examples that are to demonstrate its inherent subjectivity. The model's probability measures are independent of any particular interpretation of probability. The fact that it can subsume any predictive problem makes the information-theoretic and probabilistic setting completely general.

The further constraint of effectiveness is also both intuitively and formally extremely mild. This makes the information-theoretic, probabilistic and effective setting extremely general. However, it's not enough to refer to the uncontroversial Church-Turing Thesis in order to establish that effectiveness is in fact no restriction at all. We need a stronger version, what I call the Church-Turing-Solomonoff Thesis, that makes a claim about effectiveness of data generating processes in the world. As things stand, this claim might be false. In that case the assumption of effectiveness is a genuine loss of generality.

An answer to this threat is to interpret the class of computable measures not as environments, but as predictors. In fact, we can naturally extend this to the more consistent identification of all possible predictors with all semicomputable semimeasures. Then the assumption of an actual environment that is given by an effective measure is replaced by the assumption of a most informed predictor that is given by an effective semimeasure. The assumption that our methods of prediction must be effective, codified in the Modest Church-Turing-Solomonoff Thesis, is so natural as to function as no restriction at all.

... there are universal predictors In this universal setting, there exist predictors that adhere to Universality. There exist predictors with the property of formal universality, or dominance, a property that translates into the informal universality property of taking every possibility into consideration. Moreover, this property gives rise to what is sometimes called completeness: rapid convergence to any actual distribution. These predictors are universal in the sense that they almost always (i.e., with probability 1) converge to true predictions, whatever the actual generating distribution.

The completeness of the universal predictors is essential, and may be interpreted as showing that the universal prediction methods "work", that they are reliable. This is the content of the principle of Completeness. But, following the rejection of the method interpretation of SP, completeness by itself conveys no ideas that can further the aims of the idealized theory; and so Completeness should be rejected as a principle.

The rejection of the method interpretation also hands us a reply to what is often taken as the most serious problem of SP: the inherent subjectivity brought on by the choice of one particular member from the complete class of universal predictors. All universal predictors can be easily shown to be asymptotically equivalent, but this invariance falls short of equality. This indeterminacy within constant bounds is thematic to the subject. It appears to be an inevitable component because there seem to be no valid criteria to objectively prefer one universal predictor (universal machine, weight function) above the rest. But from the perspective of my theory interpretation, that only takes SP to state that there exists a class of universal predictors in the universal setting, there is no reason to even require to select one element from this class. The relevant thing is that this class exists; it makes no sense to worry about which predictor from this class should be selected if we can't ever put it to use anyway.

The issue of indeterminacy returns in my criticism of the alleged simplicity bias in the universal predictors. The crucial step in establishing this bias, both in the definition of algorithmic probability and of the universal mixture distribution, is the identification of simplicity with description length. While this idea is intuitively convincing, the freedom both within and in between the possible formal definitions of description length stretches this intuition to the point of breaking it. In my interpretation, indeterminacy with respect to universal predictors is not consequential; for an objective definition of simplicity, however, it is. In the end, this discussion reduces to the conceivably insoluble matter of how much of a difference is still a difference. But the fact remains that the role of simplicity in SP is certainly less obvious than generally assumed, and I take this as sufficient ground to think of Simplicity as subsidiary at best to the unproblematic and overarching principle of Universality.

If we interpret the elements of the class of semicomputable semimeasures as predictors, then the universal elements don't dominate all possible environments but all possible competing predictors. A refined convergence result then holds that the universal predictors quickly converge to the predictions of the best predictor – irrespective of the actual, possibly not even semicomputable, environment. In that case, rather than universally reliable, the universal predictors are universally optimal. We have lost all reference to the truth, but, conditional on the argument that optimality is all that we need because the predictors represent the best we can do, the replacement of the questionable assumption of effective truth by the much more plausible assumption

of effectiveness of predictors gains us full universality.

Final Remarks

My interpretation of Solomonoff Prediction is a modest one.

It takes a step back from the common interpretation that SP provides us with an idealized yet universally objective "solution to everything" method of prediction. This interpretation is either wrong or quite useless. It's also harmful to the propagation of the theory. The implausibility of this interpretation gives critics a handle to dismiss the value of the project out of hand. It gives them a reason not to take the trouble of looking further into it. It leads them to the other extreme view, that Solomonoff's theory is really nothing but a mathematical curiosity.

My interpretation takes an intermediate position. Solomonoff Prediction is more than a mathematical curiosity, because it can be convincingly brought into the context of the philosophical problem of prediction. In this thesis, I have shown how this is done.

I have argued that my interpretation is sound and best accentuates the potential value of the theory of SP. I believe that SP *is* valuable. However, a complete argument for the value of SP must include an in-depth defence of how it clarifies related philosophical issues or how it inspires practical theories. That work wasn't completed in this thesis: it wasn't its aim. Although I touched upon many of the ideas in the proximity of SP, from formal learning theory to meta-induction, I consequently stay clear of claims about the implications of this thesis for these projects. I remain silent, for instance, about the role of simplicity in MDL. Those are subjects for future investigations: I hope that the results of this thesis can provide a basis.

Bibliography

[1] H. Akaike. Information theory and an extension of the maximum likelihood principle. In B.N. Petrov and F. Csaki, editors, *Proceedings Second International Symposion on Information Theory*, pages 267–281, Budapest, 1973. Akademiai Kiado.

[2] E. Asmis. Epicurean epistemology. In K. Algra, J. Barnes, J. Mansfeld, and M. Schofield, editors, *The Cambridge History of Hellenistic Philosophy*, chapter 8, pages 260–294. Cambridge University Press, Cambridge, 1999.

[3] A. Baker. Simplicity. In E.N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, 2011.

[4] G. Barmpalias and D.L. Dowe. Universality probability of a prefix-free machine. *Philosophical Transactions of the Royal Society*, 370(1971):3488–3511, 2012.

[5] A.R. Barron, J.J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory*, 44(6):2743–2760, 1998.

[6] T. Bayes. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53:370–418, 1763.

[7] C.S. Calude. Algorithmic randomness, quantum physics, and incompleteness. In M. Margenstern, editor, *Proceedings of the Conference "Machines, Computability and Universality" (MCU 2004)*, pages 1–17, Berlin, 2005. Springer.

[8] R. Carnap. On inductive logic. *Philosophy of Science*, 12:72–97, 1945.

[9] R. Carnap. The two concepts of probability. *Philosophy and Phenomenological Research*, 5:513–532, 1945.

[10] R. Carnap. *Logical Foundations of Probability*. University of Chicago Press, Chicago, first edition, 1950.

[11] R. Carnap. *The Continuum of Inductive Methods*. University of Chicago Press, Chicago, 1952.

[12] R. Carnap. *Logical Foundations of Probability*. University of Chicago Press, second edition, 1962.

[13] R. Carnap. Replies and systematic expositions. In P.A. Schilpp, editor, *The Philosophy of Rudolf Carnap*. Open Court, 1963.

[14] N. Cesa-Bianchi and G. Lugosi. On prediction of individual sequences. *The Annals of Statistics*, 27:1865–1895, 1999.

[15] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning and Games*. Cambridge University Press, Cambridge, 2006.

[16] G.J. Chaitin. On the length of programs for computing finite binary sequences. *Journal of the Association for Computing Machinery*, 13:574–569, 1966.

[17] G.J. Chaitin. On the length of programs for computing finite binary sequences: Statistical considerations. *Journal of the ACM*, 13:547–569, 1969.

[18] G.J. Chaitin. Randomness and mathematical proof. *Scientific American*, 1975.

[19] G.J. Chaitin. A theory of program size formally identical to information theory. *Journal of the ACM*, 22:329–340, 1975.

[20] G.J. Chaitin. *The Unknowable*. Springer Verlag, Singapore, 1999.

[21] G.J. Chaitin. The limits of reason. *Scientific American*, 294(3):74–81, 2006.

[22] G.J. Chaitin. *Meta Math! The Quest for Omega*. Atlantic Books, 2006.

[23] A. Church. A set of postulates for the foundation of logic. *Annals of Mathematics*, 33:346–366, 1932.

[24] A. Church. An unsolvable problem of elementary number theory. *American Journal of Mathematics*, 58(2):345–363, 1936.

[25] A. Church. Review of Turing 1936. *Journal of Symbolic Logic*, 2:42–43, 1937.

[26] A. Church. *The Calculi of Lambda-Conversion*. Princeton University Press, Princeton, New Jersey, 1941.

[27] B.J. Copeland. Narrow versus wide mechanism,including a re-examination of Turing's views on the mind–machine issue. *Journal of Philosophy*, 97:5–32, 2000.

[28] B.J. Copeland. Accelerating Turing machines. *Minds and Machines*, 12(2):281–301, 2002.

[29] B.J. Copeland. The Church-Turing Thesis. In E.N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Fall 2008 edition, 2008.

[30] B.J. Copeland and D. Proudfoot. Alan Turing's Forgotten Ideas in Computer Science. *Scientific American*, 278(4):98–103, 1999.

[31] B.J. Copeland and R. Sylvan. Beyond the universal Turing machine. *Australasian Journal of Philosophy*, 77:46–66, 1999.

[32] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Wiley, New York, first edition, 1991.

[33] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Hoboken, New Jersey, second edition, 2006.

[34] M. Curd and J.A. Cover. *Philosophy of Science - The Central Issues*. W.W. Norton and Company, New York, NY, 1998.

[35] E.B. Davies. Building infinite machines. *British Journal for the Philosophy of Science*, 52(4):671–682, 2001.

[36] M.D. Davis. The myth of hypercomputation. In C. Teuscher, editor, *Alan Turing: The Life and Legacy of a Great Thinker*, pages 195–212. Springer, Berlin, 2006.

[37] M.D. Davis. Why there is no such discipline as hypercomputation. *Applied Mathematics and Computation*, 178:4–7, 2006.

[38] A.P. Dawid. Present position and potential developments: Some personal views. Statistical theory: The prequential approach. *Journal of the Royal Statistical Society. Series A (General)*, 147(2):278–292, 1984. The 150th Anniversary of the Royal Statistical Society.

[39] A.R. Day. Increasing the gap between descriptional complexity and algorithmic probability. *Transactions of the American Mathematical Society*, 363(10):5577–5604, 2011.

[40] B. de Finetti. La prévision: Ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, 17, 1937.

[41] S. de Rooij and P. D. Grünwald. Luckiness and regret in minimum description length inference. In P.S. Bandyopadhyay and M. Forster, editors, *Philosophy of Statistics*, volume 7 of *Handbook of the Philosophy of Science*. Elsevier, 2011.

[42] R.M. de Wolf. Philosophical applications of computational learning theory: Chomskyan innateness and Occam's razor. Master's thesis, Erasmus Universiteit Rotterdam, 1997.

[43] J.-P. Delahaye. Randomness, unpredictability and absence of order. In J. Dubucs, editor, *Philosophy of Probability*, pages 145–167. Kluwer Academic Publishers, Dordrecht, 1993.

[44] D.E. Deutsch. Quantum theory, the Church-Turing principle and the universal quantum computer. *Proceedings of the Royal Society of London A*, 400:97–117, 1985.

[45] D.L. Dowe. MML, hybrid Bayesian network graphical models, statistical consistency, invariance and uniqueness. In *Handbook of the Philosophy of Science, Volume 7: Handbook of Philosophy of Statistics*, pages 901–982. Elsevier, 2011.

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

[47] J. Earman. *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*. Bradford, 1992.

[48] R.L. Ellis. Remarks on an alleged proof of the method of least squares, contained in a late number of the *Edinburgh Review*. In W. Walton, editor, *Mathematical and Other Writings of R.L. Ellis*, pages 53–61. Cambridge University Press, Cambridge, 1863.

[49] M. Feder. Maximum entropy as a special case of the minimum description length criterion. *IEEE Transactions on Information Theory*, 32(6):847–849, 1986.

[50] W. Feller. *An introduction to probability theory and its applications. Vol. II.* John Wiley & Sons Inc., New York, 1971.

[51] E. Fredkin. Digital mechanics: An information process based on reversible universal cellular automata. *Physica D*, 45:254–270, 1990.

[52] E. Fredkin. An introduction to digital philosophy. *International Journal of Theoretical Physics*, 42(3):189–247, 2003.

[53] P. Gács. On the relation between descriptional complexity and algorithmic probability. *Theoretical Computer Science*, 22:71–93, 1983.

[54] P. Gács. Review of Gregory J. Chaitin, Algorithmic Information Theory. *Journal of Symbolic Logic*, 54:624–627, 1989.

[55] P. Gács. Expanded and improved proof of the relation between description complexity and algorithmic probability. Unpublished manuscript, 2008.

[56] H. Gaifman and M. Snir. Probabilities over rich languages, testing and randomness. *Journal of Symbolic Logic*, 47(3):495–548, 1982.

[57] M.C. Galavotti. *Philosophical Introduction to Probability*. CSLI Publications, Stanford, 2005.

[58] R. Gandy. Church's thesis and principles for mechanisms. In J. Barwise, H.J. Keisler, and K. Kunen, editors, *The Kleene Symposium*. North-Holland, 1980.

[59] R. Gandy. The confluence of ideas in 1936. In R. Herken, editor, *The Universal Turing Machine: A Half-Century Survey*, pages 51–102. Oxford University Press, Oxford, 1988.

[60] C. Glymour. Why I am not a Bayesian. *Theory and evidence*, pages 63–93, 1981.

[61] K. Gödel. On undecidable propositions of formal mathematical systems. Lecture notes taken by Kleene and Rosser at the Institute for Advanced Study, 1934. In M.D. Davis, editor, *The undecidable. Basic papers on undecidable propositions, unsolvable problems and computable functions*, pages 39–74. Raven Press, Hewlet, New York, 1965.

[62] K. Gödel. Remarks before the 1946 Princeton Bicentennial Conference on Problems in Mathematics. In M.D. Davis, editor, *The undecidable. Basic papers on undecidable propositions, unsolvable problems and computable functions*, pages 84–88. Raven Press, Hewlet, New York, 1965.

[63] E.M. Gold. Language identification in the limit. *Information and Control*, 10:447–474, 1967.

[64] N. Goodman. *Fact, Fiction, & Forecast*. Harvard University Press, Cambridge, Massachusetts, 1955.

[65] P.D. Grünwald. Maximum entropy and the glasses you are looking through. In *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence (UAI 2000)*, pages 238–246, 2000.

[66] P.D. Grünwald. *The Minimum Description Length Principle*. MIT Press Books. MIT Press, 2007.

[67] P.D. Grünwald and P.M.B. Vitányi. Algorithmic information theory. In P. Adriaans and J.F.A.K. van Benthem, editors, *Handbook of the Philosophy of Science, Vol. 8: Philosophy of Information*, pages 289–325. Elsevier Science, 2008.

[68] A. Hájek. Interpretations of probability. In E.N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Summer 2012 edition, 2012.

[69] J. Herbrand. Sur la non-contradiction de l'arithmetique. *Journal für die reine und angewandte Mathematik*, 166:1–8, 1932.

[70] M. Hesse. *The Structure of Scientific Inference*. Cambridge: Cambridge University Press, 1974.

[71] A. Hodges. Alan Turing. In E.N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Summer 2011 edition, 2011.

[72] P. Horwich. Wittgensteinian bayesianism. *Midwest Studies in Philosophy*, 18(1):62–75, 1993.

[73] D. Huffman. A method for the construction of minimum-redundancy codes. *Proceedings of the Institute of Radio Engineers*, 40(9):1098–1101, 1952.

[74] D. Hume. *A Treatise of Human Nature*. Clarendon Press, Oxford, 1888. Originally published 1739–40.

[75] M. Hutter. A theory of universal artificial intelligence based on algorithmic complexity. Technical Report cs.AI/0004001, München, 62 pages, 2000.

[76] M. Hutter. Convergence and error bounds for universal prediction of nonbinary sequences. In *Proceedings of the 12th European Conference on Machine Learning (ECML-2001)*, volume 2167 of *Lecture Notes in Artificial Intelligence*, pages 239–250, Freiburg, 2001. Springer, Berlin.

[77] M. Hutter. New error bounds for Solomonoff prediction. *Journal of Computer and System Sciences*, 62(4):653–667, 2001.

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

[79] M. Hutter. Optimality of universal Bayesian prediction for general loss and alphabet. *Journal of Machine Learning Research*, 4:971–1000, 2003.

[80] M. Hutter. Sequence prediction based on monotone complexity. In B. Schölkopf and M.K. Warmuth, editors, *Proceedings 16th Annual Conference on Learning Theory (COLT'03)*, volume 2777 of *Lecture Notes in Artificial Intelligence*, pages 506–521, Washington, DC, 2003. Springer, Berlin.

[81] M. Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Springer, Berlin, 2005.

[82] M. Hutter. Sequential predictions based on algorithmic complexity. *Journal of Computer and System Sciences*, 72(1):95–117, 2006.

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

[84] M. Hutter. Open problems in universal induction & intelligence. *Algorithms*, 3(2):879–906, 2009.

[85] M. Hutter. One decade of universal artificial intelligence. In *Theoretical Foundations of Artificial General Intelligence*. Atlantis Press, 2012. To appear.

[86] M. Hutter and An. A. Muchnik. Universal convergence of semimeasures on individual random sequences. In *Proceedings of the 15th International Conference on Algorithmic Learning Theory (ALT'04)*, volume 3244 of *Lecture Notes in Artificial Intelligence*, pages 234–248, Padova, 2004. Springer, Berlin.

[87] E.T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 108(2):171–190, 1957.

[88] E.T. Jaynes. Information theory and statistical mechanics i. *Physical Review*, 106(4):620–630, 1957.

[89] E.T. Jaynes. Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics*, 4(3):227 – 241, 1968.

[90] R.C. Jeffrey, editor. *Studies in Inductive Logic and Probability*, volume II. University of California Press, Berkeley and Los Angeles, 1980.

[91] H. Jeffreys. *Theory of Probability*. Oxford: Oxford University Press, third edition, 1967.

[92] H. Jeffreys. *Theory of Probability*. Clarendon Press, Oxford, 1985.

[93] K.T. Kelly. *The Logic of Reliable Inquiry (Logic and Computation in Philosophy)*. Oxford University Press, USA, 1996.

[94] K.T. Kelly. Uncomputability: the problem of induction internalized. *Theoretical Compututer Science*, 317(1-3):227–249, 2004.

[95] K.T. Kelly. Ockham's razor, truth, and information. In J.F.A.K. van Benthem and P. Adriaans, editors, *Handbook of the Philosophy of Information*, pages 321–360. Elsevier, Dordrecht, 2008.

[96] K.T. Kelly and C. Glymour. Why probability does not capture the logic of scientific justification. In C. Hitchcock, editor, *Contemporary Debates in the Philosophy of Science*, pages 94–114. London: Blackwell, 2004.

[97] K.T. Kelly, O. Schulte, and C. Juhl. Learning theory and the philosophy of science. *Philosophy of Science*, 64(2):245–267, 1997.

[98] J.G. Kemeny. The use of simplicity in induction. *Philosophical Review*, 62(3):391–408, 1953.

[99] J.G. Kemeny. Fair bets and inductive probabilities. *Journal of Symbolic Logic*, 20:263–273, 1955.

[100] J.M. Keynes. *A Treatise on Probability*. Macmillan, London, 1921.

[101] S.C. Kleene. General recursive functions of natural numbers. *Mathematische Annalen*, 112:727–742, 1936.

[102] S.C. Kleene. Recursive predicates and quantifiers. *Transactions of the American Mathematical Society*, 53(1):41–73, 1943.

[103] S.C. Kleene. *Introduction to metamathematics*. North-Holland, 1952.

[104] A.N. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer, Berlin, 1933.

[105] A.N. Kolmogorov. Three approaches to the definition of the concept "quantity of information". *Problemy Peredači Informacii*, 1(vyp. 1):3–11, 1965.

[106] A.N. Kolmogorov. Complexity of algorithms and objective definition of randomness. *Uspekhi Matematicheskikh Nauk*, 29(4):155, 1974. Abstract of a talk at the Moscow Mathematical Society meeting of April 4, 1974. In Russian.

[107] A.N. Kolmogorov. Talk at the Information Theory Symposion in Talinn, Estionia, 1974.

[108] L.G. Kraft. A device for quantizing, grouping, and coding amplitude modulated pulses. M.Sc. Thesis, Dept. of Electrical Engineering, MIT, Cambridge, Mass., 1949.

[109] T.S. Kuhn. *The structure of scientific revolutions*. University of Chicago Press, Chicago, second edition, 1970.

[110] T. Lattimore and M. Hutter. On Martin-Löf convergence of Solomonoff's mixture. To appear.

[111] S. Legg. Solomonoff induction. Technical Report CDMTCS-030, Centre for Discrete Mathematics and Theoretical Computer Science. University of Auckland, 1997.

[112] S. Legg. *Machine Super Intelligence*. PhD thesis, Università della Svizzera Italiana, 2008.

[113] S. Legg and M. Hutter. Universal intelligence: A definition of machine intelligence. *Minds & Machines*, 17(4):391–444, 2007.

[114] L.A. Levin. On the notion of random sequence. *Soviet Mathematics – Doklady*, 14:1413–1416, 1973.

[115] L.A. Levin. Laws of information conservation (nongrowth) and aspects of the foundation of probability theory. *Problems of Information Transmission*, 10:206–210, 1974.

[116] L.A. Levin. Randomness conservation inequalities; information and independence in mathematical theories. *Information and Control*, 61(1):15–37, 1984.

[117] M. Li and P.M.B. Vitányi. Inductive reasoning and Kolmogorov complexity. *Journal of Computer and System Sciences*, 44(2):343–384, 1992.

[118] M. Li and P.M.B. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer Verlag, New York, first edition, 1993.

[119] 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.

[120] P. Lipton. *Inference to the Best Explanation*. Routledge, New York, NY, 1991.

[121] P. Martin-Löf. The definition of random sequences. *Information and Control*, 9:602–619, 1966.

[122] M.L. Minsky. Steps towards artificial intelligence. In *Proceedings of the IRE*, pages 8–30, 1961.

[123] M.L. Minsky. Problems of formulation for artificial intelligence. In R.E. Bellman, editor, *Mathematical Problems in the Biological Sciences*, Proceedings of Symposia in Applied Mathematics XIV, page 43, Providence, Rhode Island, 1962. American Mathematical Society.

[124] M.L. Minsky. *Computation: Finite and Infinite Machines*. Prentice-Hall, Inc., Upper Saddle River, New Jersey, 1967.

[125] M. Müller. Stationary algorithmic probability. *Theoretical Computer Science*, 411(1):113–130, 2010.

[126] E. Nagel. Carnap's theory of induction. In P. A. Schilpp, editor, *The philosophy of Rudolf Carnap*, pages 785–826. Open Court, 1963.

[127] A. Nies. *Computability and randomness*. Oxford logic guides. Oxford University Press, 2009.

[128] P. Odifreddi. *Classical recursion theory: the theory of functions and sets of natural numbers*. North-Holland, 1989.

[129] G. Piccinini. Computation in physical systems. In E.N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Fall 2012 edition, 2012.

[130] G. Piccinini and A. Scarantino. Information processing, computation, and cognition. *Journal of Biological Physics*, 37(1):1–38, 2011.

[131] I. Pitowsky. The physical Church-Turing thesis and physical computational complexity. *Iyyun*, 39:81–99, 1990.

[132] M.B. Pour-El and I. Richards. A computable ordinary differential equation which possesses no computable solution. *Annals of Mathematical Logic*, 17:61–90, 1979.

[133] M.B. Pour-El and I. Richards. The wave equation with computable initial data such that its unique solution is not computable. *Advances in Mathematics*, 39(3):215–239, 1981.

[134] H. Putnam. Degree of confirmation and inductive logic. In P.A. Schilpp, editor, *The Philosophy of Rudolph Carnap*, pages 761–783. Open Court, 1963.

[135] H. Putnam. Probability and confirmation. In S. Morgenbesser, editor, *Philosophy of Science Today*. Basic Books, New York, 1967.

[136] H. Putnam. *Representation and Reality*. MIT Press, Cambridge, Massachusetts, 1988.

[137] P. Raatikainen. Complexity and information – a critical evaluation of algorithmic information theory. Technical Report 2, Department of Philosophy, University of Helsinki, 1998.

[138] P. Raatikainen. Algorithmic information theory and undecidability. *Synthese*, 123:217–225, 2000.

[139] F.P. Ramsey. Truth and Probability. In R. B. Braithwaite, editor, *The Foundations of Mathematics and other Logical Essays*, chapter 7, pages 156–198. Brace & Co., 1926.

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

[141] H. Reichenbach. *Wahrscheinlichkeitslehre: eine Untersuchung über die logischen und mathematischen Grundlagen der Wahrscheinlichkeitsrechnung*. Sijthoff, Leiden, 1935.

[142] J.J. Rissanen. Modeling by shortest data description. *Automatica*, 14:445–471, 1978.

[143] J.J. Rissanen. A universal prior for integers and estimation by minimum description length. *Annals of Statistics*, 11:416–431, 1982.

[144] J.J. Rissanen. *Stochastic Complexity in Statistical Inquiry*. World Scientific, Singapore, 1989.

[145] H. Rogers. *Theory of Recursive Functions and Effective Computability*. MIT Press, Cambridge, Massachusetts, second edition, 1987.

[146] R. Rosenkrantz. Simplicity. In W. Harper and C. Hooker, editors, *Foundations and Philosophy of Statistical Inference*, pages 167–196. Boston: Reidel, 1976.

[147] W.C. Salmon. *The Foundations of Scientific Inference*. University of Pittsburgh Press, 1967.

[148] W.C. Salmon. Rationality and objectivity in science or Tom Kuhn meets Tom Bayes. In C. W. Savage, editor, *Scientific Theories*, volume XIV of *Minnesota Studies in the Philosophy of Science*. University of Minnesota Press, Minneapolis, MN, 1990.

[149] L.J. Savage. *The Foundations of Statistics*. John Wiley and Sons, New York, 1954.

[150] C.-P. Schnorr. Process complexity and effective random tests. *Journal of Computer and System Sciences*, 7:376–388, 1973. Fourth Annual ACM Symposium on the Theory of Computing (Denver, Colorado, 1972).

[151] O. Schulte. Means-ends epistemology. *The British Journal for the Philosophy of Science*, 50:1–31, 1999.

[152] G. Schurz. The meta-inductivist's winning strategy in the prediction game: A new approach to Hume's problem. *Philosophy of Science*, 75:278–305, 2008.

[153] G. Schurz. Local, general and universal prediction strategies: A game-theoretical approach to the problem of induction. In M. Suárez, M. Dorato, and M. Rédei, editors, *EPSA Epistemology and Methodology of Science: Launch of the European Philosophy of Science Association*, pages 267–278, Dordrecht, 2010. Springer.

[154] J. Searle. *The Rediscovery of the Mind*. MIT Press, Cambridge, Mass., 1992.

[155] C.E. Shannon. The mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948.

[156] A. Shimony. Scientific inference. In *Pittsburgh Studies in the Philosophy of Science, vol. 4.*, pages 79–172. University of Pittsburgh Press, 1970.

[157] W. Sieg. On computability. In A. Irvine, editor, *Philosophy of Mathematics (Handbook of the Philosophy of Science)*, pages 535–630. Elsevier, Amsterdam, 2006.

[158] H.T. Siegelmann. Computation beyond the Turing limit. *Science*, 268:545–548, 1995.

[159] H.T. Siegelmann. *Neural Networks and Analog Computation*. Birkhäuser, Boston, 1999.

[160] R.I. Soare. *Computability theory and its applications*. Springer-Verlag, Heidelberg, 2012. To appear.

[161] R.J. Solomonoff. An inductive inference machine. In *IRE Convention Record, Section on Information Theory, Part 2*, pages 56–62, 1957.

[162] R.J. Solomonoff. A preliminary report on a general theory of inductive inference. (revision of Report V-131). Technical Report ZTB-138, Zator Co. and Air Force Office of Scientific Research, Cambridge, Mass., Nov 1960.

[163] R.J. Solomonoff. A formal theory of inductive inference. Part I. *Information and Control*, 7:1–22, 1964.

[164] R.J. Solomonoff. A formal theory of inductive inference. Part II. *Information and Control*, 7:224–254, 1964.

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

[166] R.J. Solomonoff. *The Time Scale of Artificial Intelligence: Reflections on Social Effects*, volume 5 of *Human Systems Management*, pages 149–153. Elsevier Science Publishers, North-Holland, 1985.

[167] R.J. Solomonoff. The application of algorithmic probability to problems in artificial intelligence. In *Uncertainty in Artificial Intelligence, Kanal, L.N. and Lemmer, J.F. (Eds), Elsevier Science Publishers B.V*, pages 473–491. Elsevier Science Publishers, 1986.

[168] R.J. Solomonoff. Does algorithmic probability solve the problem of induction? In *Proceedings of the Information, Statistics and Induction in Science Conference*, 1996.

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

[170] R.J. Solomonoff. The universal distribution and machine learning. *The Computer Journal*, 46(3):598–601, 2003. The Kolmogorov Lecture, Feb. 27, 2003, Royal Holloway, University of London.

[171] R.J. Solomonoff. The probability of "undefined" (non-converging) output in generating the universal probability distribution. *Information Processing Letters*, 106(6):238–240, 2008.

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

[173] A.M. Turing. Systems of logic based on ordinals. *Proceedings of the London Mathematical Society*, pages 161–228, 1939.

[174] J.B.M. Uffink. Grondslagen van het waarschijnlijkheidsbegrip, 1990. Manuscript.

[175] J.B.M. Uffink. Can the maximum entropy principle by explained as a consistency requirement? *Studies in History and Philosophy of Modern Physics*, 26(3):223–261, 1995.

[176] A.P.M. van den Bosch. Computing simplicity: About the role of simplicity in discovery, explanation, and prediction. Master's thesis, Rijksuniversiteit Groningen, 1994.

[177] M. van Lambalgen. *Random sequences*. PhD Dissertation, Universiteit van Amsterdam, The Netherlands, 1987.

[178] M. van Lambalgen. Algorithmic information theory. *Journal of Symbolic Logic*, 54:1389–1400, 1989.

[179] N.K. Vereshchagin and P.M.B. Vitányi. Kolmogorov's structure functions and an application to the foundations of model selection. *IEEE Transactions on Information Theory*, 50(12):3265–3290, 2004.

[180] J. Vickers. The problem of induction. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Winter 2012 edition, 2012.

[181] P.M.B. Vitányi. Algorithmic statistics and Kolmogorov's structure functions. In P.D. Grünwald, I.J. Myung, and M.A. Pitt, editors, *Advances in Minimum Description Length: Theory and Applications*, pages 151–174. MIT Press, 2005.

[182] V. Vovk. Aggregating strategies. In *Proceedings Third ACM Workshop on Computational Learning Theory (COLT 1990)*, pages 371–383, 1990.

[183] C.S. Wallace and D.M. Boulton. An information measure for classification. *Computer Journal*, 11(2):185–194, 1968.

[184] J.A. Wheeler. Information, physics, quantum: The search for links. In W. Zurek, editor, *Complexity, Entropy, and the Physics of Information*. Addison-Wesley, Redwood City, California, 1990.

[185] D.G. Willis. Computational complexity and probability constructions. *Journal of the Association for Computing Machinery*, 17(2):241–259, 1970.

[186] L. Wittgenstein. *Remarks on the Philosophy of Psychology*, volume 1. Blackwell, Oxford, 1980.

[187] S. Wolfram. *A New Kind of Science*. Wolfram Media, Champaign, Illinois, 2002.

[188] I. Wood, P. Sunehag, and M. Hutter. (Non-)equivalence of universal priors. In *Proceedings Solomonoff 85th Memorial Conference*, LNAI, Melbourne, Australia, 2011. Springer.

[189] S.L. Zabell. Carnap and the logic of inductive inference. In D.M. Gabbay, S. Hartmann, and J. Woods, editors, *Handbook of the History of Logic. Volume 10: Inductive Logic.*, pages 265–309. Elsevier BV, 2009.

[190] K. Zuse. *Rechnender Raum*. Friedrich Vieweg & Sohn, Braunschweig, 1969.

[191] 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 Math. Surveys*, page 11, 1970.

Symbol Index

Name Index

Index