The philosophy of learning: the cooperative computational universe

Pieter Adriaans
Human-Computer Studies
University of Amsterdam
Kruislaan 419
Amsterdam
pietera@science.uva.nl.

Abstract

In this paper I discuss various philosophical issues in relation to the formal study of learning as data compression, also known as Minimum Description Length or Two Part Code optimization. I show that learning is intimately related to basic questions in epistemology. Central is the problem of the efficiency of human learning. Drawing on fundamental insights from complexity theory, information theory and thermodynamics I sketch a unifying view that clarifies this human efficiency: a universe in which we can compute is necessarily a cooperative universe in the sense that it produces data sets that can easily be compressed.

1 Introduction

In the summer of 1956 a number of scientists gathered at the Dartmouth College in Hanover, New Hampshire. Their goal was to study human intelligence with the help of computers. Their central hypothesis was: "that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it." On that conference, where amongst others John McCarthy, Claude Shannon and Marvin Minsky were present, the new discipline of Artificial Intelligence was born. It is striking 'learning' was considered to be an important aspect of human intelligence from the start. A better understanding of the phenomenon of learning was high on the agenda of the emerging young science.

Now, fifty years later, the study of learning is one of the success stories of AI. There is a multitude of learning techniques for the computer. Data mining techniques are being used for marketing, stock management, production optimization and fraud detection in the commercial domain. Biologically inspired learning models such as neural networks and genetic algorithms are being used to simulate human cognition and evolution. In disciplines like computer vision and computational linguistics machine learning is in the center of interest (Kearns and Vazirani [1994], Mitchell [1997], Adriaans and Zantinge [1997], Cornuéjols and Miclet [2003]).

But, researchers do not have much reason to sit back and rest, because there is still a whole list of questions that are begging for answers. One of the biggest embarrassments is that we still do not know what learning is exactly. The toolbox of a machine learner looks like a haphazardly collected bunch of

screwdrivers, hammers en chisels of dubious origin. For some jobs they work, but we do not understand why, for others they do not work and we also do not understand why. One thing is certain. There will never be a general theory that explains what learning exactly is.

Philosophy of information

It is clear that with the adventure of artificial intelligence we have hit upon a problem domain that has much wider repercussions than the creation of intelligent computers. Recently a new discipline has emerged: the philosophy of information (Floridi [2004]).[1] This discipline reformulates central questions of philosophy from the perspective of modern insights from computer science. Developments like these, urge us to formulate the question of the relation between philosophy on one side and logic, mathematics, theory of information and computation on the other.

First of all philosophy is not science. It takes a meta-position and is always at most *a reflection on science* and scientific results. It is not the primary task of the philosopher to formulate and prove theorems. It is his task to reflect on the consequences of theorems and theories. On the other hand *philosophy can not claim to have any form of privileged access to reality.* There is no fixed archimedean position from which the philosopher can judge the results of scientific endeavors. [2] Philosophy and science therefore are doomed to live permanently in each others shadow without any possibility of a final reconciliation. Any scientific result can be made object of philosophical analysis, but ... only, or predominantly, in terms of the concepts that the sciences have constructed themselves. Philosophy therefore is at its best when it is in dialogue with foundational programs of science and the humanities. The more it removes itself from these central issues, the more substance it loses and the more it deteriorates in to a (possibly brilliant) literary exercise at best. In this sense philosophical reflection may be seen as an inherent and necessary aspect of scientific heuristics. It provides us with a rich historical context of 2500 years of reflection on foundational programs and invites us to investigate the more extreme consequences of our theories and models.

The study of theory of knowledge, theory of information and computation, methodology of science, theory of induction and meta-mathematics share a common history in which related questions have been analyzed in different guises. The work of Solomonoff and Kolmogorov provides direct answers to questions about the nature of knowledge and induction proposed by Carnap and the Wiener Kreis and much earlier Kant and Hume. In this light one has to interpret the reflections on theory of information and learning I present below.

[1] See the chapter by Floridi in this book

[2] Specifically: no privileged direct access to ones own consciousness, no Husserlian epoche, no historical laws of materialism, no recourse to immediately given sense data, no special rapport with Being itself, etc.

Philosophy of learning

First I show that the question of the essence of learning is embedded in fundamental epistemological questions. The old philosophical problem of the essence of knowledge is fundamentally associated with learning. The notion of efficiency of learning plays an essential role in this context. Our models of learning show us that tasks, like learning a language, that human beings perform without too much problems, are from a formal point of view extremely complex and next to impossible. This leaves us with the riddle of human efficiency. I show how the contours of an analysis of this mysterious efficiency of human learning gets shape in the light of recent insights from complexity theory and thermodynamics. Central questions in this respect are:

Question 1.1 *What is learning?*

Question 1.2 *What are data sets from which we can learn?*

Question 1.3 *What kind of systems produce those data sets?*

The answer to the first question is: learning is algorithmic compression of data sets. Not all forms of learning are caught by this definition, but a broad class of philosophically relevant learning phenomena fall under this description. [3] The answer to the second question is: data sets that can be compressed by a computer algorithm without too much effort. [4] An answer to the third question is - quite naturally - systems with relative low entropy: i.e. self-organizing systems, systems that are not in a state of thermal equilibrium and systems that redirect energy from their environment in order to keep their internal entropy lower than that of the environment. This kind of self-organization is typical for life and for computational processes. The picture that emerges is that those systems in nature that produce data sets from which something can be learned are by necessity systems with a relatively low entropy. The data sets themselves consequently have low entropy and are easy to decipher. This seems to be the solution to the problem of the efficiency of our learning algorithms. A deep analysis of the idea that the universe can be interpreted as a computational process shows that nature necessarily acts as a cooperative teacher. This is a philosophical insight that transcends the local context of Artificial Intelligence. At the same time these insights help us to develop new algorithms that solve problems from every day life. Learning as data compression helps us to classify viruses, analyze music (Cilibrasi and Vitanyi [2005]) and to learn languages (Adriaans [2001]).

[3] Neural networks, genetic algorithms, decision tree induction, clustering, nearest neighbor, support vector machines, association rules, to name a few. As a counter example: simple rote learning of a finite set of facts does not necessarily involve compression of data.

[4] Technically: data sets that can be compressed by means of constructive resource bounded compression. The 'without too much effort' restriction is added because it actually is possible to construct highly compressible data sets that from the outside look random, e.g. encrypted data or expansions of very special real numbers like π and e. There are no general algorithms to compress these sets. It is highly unlikely that these data sets occur frequently in nature. Anyhow, we would not notice them.

A short historical digression

The notion that knowing something implied knowing its 'form' goes back to Plato's theory of ideas as forms. Aristotle's more empirical doctrine of the four causes (causalis, finalis, formalis and efficiens) also distinguishes the notion of form as a crucial element of knowledge. The original technical notion of the Latin word 'in-formare' (giving form to something, impressing ideas/forms in the mind in the Platonic sense) that is found in the writings of Cicero[5] and Augustine seems to have played no role in the emergence of the modern concept of information. The word 'idea' seems the true modern heir of the classical term 'information' (Capurro [1978], Capurro and Hjørland [2003]).

In the 15th century the French term 'information' finds its way into the colloquial vocabulary of European languages with various subtle differences in meaning, clustering around meanings like 'investigation', 'education', 'the act of informing or communicating knowledge', 'intelligence' etc. After Descartes the technical term seems to vanish from the philosophical debate. It does not play any specific role in the work of a broad philosopher like Kant. There is no lemma on information in Windelbands famous 'Lehrbuch der Geschichte der Philosophie' from 1889 (Windelband [1921]). Even Edward's Encyclopedia of Philosophy from 1967 does not have a separate lemma on information (Edwards [1967]). The same holds for the well-known History of Logic written by Kneale and Kneale that first appeared in 1962 (Kneale and Kneale [1988]). In short the term 'information' seems to have been absent from the philosophical dialogue for hundreds of years.

In the history of philosophy the phenomenon of learning for a long time only has been studied implicitly, because it is related to knowledge, but since circa 1700 AD the problem of learning is placed explicitly on the philosophical agenda. A key insight in the study of the history of the concept of information is formulated in this book by Devlin and Rosenberg in their chapter on information in the social sciences. The basic idea is that information is an abstract notion that is the natural byproduct of the emergence of modern media. When human communication was transformed from a direct dialogical interaction between individuals to an interaction that was mediated by technology (telescopes, microscopes, books, newspapers, the telephone, television, internet etc.) the need to create an abstract umbrella term to denote the 'stuff' that was flowing between sender and receiver of a message emerged. In this respect the emergence of the empirical sciences in the 17th century is a central period in history of the conceptualization of information.

Descartes (1596-1650) formulated a firm mathematical framework for the description of the material world, but his dualism prevented him from understanding the interplay between language and the growth of knowledge. For

[5] Cicero used the word information as a translation of the Epicurean notion of 'prolepsis', i.e. a representation in the mind. A notion that can be compared to the later use of the word 'idea' by Descartes and Locke. See 'On the nature of the Gods", I, 43. Also Greek terms like 'hypothesis' and 'eidos' were translated with the term 'information' by Latin authors (Capurro [1978]).

Descartes, man's rationality was equivalent to mastering language and was an innate quality. The communication between the res extensa and the res cogitans remained a central problem. Descartes is important because he is the first philosopher who formulated a theoretical framework in which the mediation between mind and body, between the knower and the known becomes problematic. With hindsight one could say that in the work of Descartes the need for *an abstract concept of mediation between knower and the known*, i.e. a concept of information, is identified for the first time. Descartes' metaphysics can not describe such a mediation. Because of this lack, he was incapable of developing an adequate philosophical theory of language and thus of an adequate conceptualization of the interplay between language and knowledge.

The next philosopher to take up this challenge was Locke (1632-1704) who developed a psychological version of carthesian dualism in the "Essay concerning human understanding" (1690)(Locke [1961]). The cartesian cogito becomes a epistemological subject that starts as a tabula rasa and is gradually filled up with 'ideas' that find their origin in experience. Descartes had formulated the notion of ideas as innate forms of thought but Locke is quite liberal in his concept of an 'idea': "*whatsoever is the object of understanding when a man thinks ... whatever is meant by phantasm, notion, species, or whatever it is which the mind can be employed about when thinking*". (Essay, I,i,8) This abstract notion of an idea, as a qualitative building block of knowledge, can be interpreted as a philosophical precursor of the modern concept of information. Ideas flow from the knower to the known, they can be isolated and combined in to new knowledge. When we receive ideas our knowledge grows.

This conceptualization of the growth of knowledge in terms of the combination of 'chunks' of knowledge implied a reformulation of a number of central problems in philosophy that would dominate the discussion for the next centuries. Central questions are:

- Can we validate general statements about the properties of a class on the basis of a finite number of observations of members of that class? Can we derive the statement "All swans are white" on the basis of "All swans we have seen so far are white"?
- Can we generalize from the past to the future?
- What part of knowledge is a priori, what part a posteriori?

In *An Enquiry Concerning Human Understanding*, (par. 4.1.20-27, par. 4.2.28-33) the philosopher Hume (1711-1776) argued that there is no logical necessity that the future will resemble the past. The insight that it is impossible to select the best theory to explain a set of observations with absolute certainty, is known as the induction problem since Hume (Hume [1914]). It denies science the possibility to formulate universal laws with absolute certainty. Several philosophers have tried to deal with this problem. It was the main motivation for the development of Kant's transcendental philosophy in the Kritik der reinen Vernuft. Kant's attempt is the last major effort to bridge the gap

between empirical science and traditional philosophy striving at the formulation of absolute truths. The empiricist program was revived by the so-called Vienna circle in the beginning of the 20th century. The ambition was to seek the foundation of science in the analysis of elementary phenomena that could be observed empirically. Needless to say that with this methodology the induction problem is a major obstacle for science. Popper, who occasionally attended meetings of the Vienna circle, formulated a solution in terms of the asymmetry between verification and falsification (Popper [1952]). Although this solved part of the problem the issue of heuristics remained open (Context of discovery versus context of justification). One solution to the induction problem is to view scientific knowledge as being essentially statistical. The concept of probability is far from harmless from a philosophical point of view (Hájek [2002]). Carnap [1950] has argued that there exist two very distinct forms of probability: a priori probability or "Rational credibility" and empirical probability in the sense of "limiting relative frequency of occurrence". Indeed there seems to be a distinct difference between the use of the notion of probability in observations like: "It is highly probable that an English sentence contains more e's than q's" and "It is highly probable that life on earth originated from outer space". The first is a statement about the frequency of letters in English. It can be corroborated by a sequence of experiments. The second statement seems different. It has prima facie nothing to do with limiting frequency. It can not be corroborated by experiments. Even if our planet was the only planet in the universe with life, the statement still could be true. It seems to express a rational belief that somebody could have after carefully examining the evidence. Black [1967] has criticized Carnap: different modes of verification for probability statements do not imply that there necessarily exist different notions of probability. The fact remains that we sometimes make judgements about the probability of individual isolated structures. This seems to involve a notion of a priori probability. If we can assign a priori probabilities to theories and data sets and conditional probabilities to a data set given a theory, then we can calculate the probability of a theory given a data set. The formulation of an exact answer to these theoretical questions is one of the great achievements of computer science in the 20th century. Solomonoff defined the idea of algorithmic complexity of a binary object as the shortest program that computes this object on a universal reference Turing machine (Solomonoff [1997]).[6]. He showed that the algorithmic or Kolmogorov complexity of an object is associated with an a priori probability of this object. It allows us in theory to assign an a priori probability as well as a complexity to an individual binary object. (universal distribution). This is the basis for modern theories about learnability and studies of methodology of science.

A central concept that ties information theory and learning together is the so-called Minimum Description Length Principle (MDL)(Rissanen [1999]). Below I will give a formal treatment of the principle but the main idea is that formal

[6] The same concept was somewhat later discovered independently by Kolmogorov and Chaitin.

representations of scientific theories can be used to compress data sets with empirical observations. The shortest adequate MDL code explaining a data set will be the one that minimizes the sum of a description, in bits, of the theory, plus a description, in bits, of the set of observations given the theory. One could think of the observations of Tycho Brahe and Keppler's laws as theory. The laws of Keppler explain the observations of Tycho Brahe, because these observations can be represented concisely using these laws. One of the main ambitions of this paper is to study the philosophical implications of this concept. The theory of Kolmogorov complexity provides us with an excellent framework for a philosophical analysis of the concepts behind MDL. This is, in my view, the form in which the problem of induction should be studied in the current context of philosophy of information.

The MDL principle is often described as being equivalent to Ockham's razor (entia non sunt multiplicanda preater necessitate, William of Ockham, ca. 1290-1349). An association that is debatable, since Ockham's razor is related to a specific nominalistic critique of Plato's theory of ideas (as defended by Duns Scotus, 1266-1308) that is quite far removed from the general problem of induction. In fact the idea of explaining a certain set of observations in terms of an optimized two-part code (Theory + Data encoded with the theory) could as well be interpreted as a Platonic ambition, where the Theory is the *ideal* description of the data and the Data encoded with the theory is a description of the noise, or *faults*, in the data. The underlying problem seems to have a different nature: the question of the regularity of nature, or in other words the notion of a cooperative universe.

2 An uneasy marriage between learning and knowing: participation versus construction

A theory of learning has consequences in at least three areas:

- Theory of knowledge: how do we gather knowledge?
- Cognition: how does our brain work?
- Methodology of science: how do we construct scientific knowledge?

Knowledge and learning always have had a bit uneasy relationship in philosophy. The subject easily could fill a book in itself. A clear picture emerges if we try to develop a simple logic of learning and knowing. We can adopt two axioms:

1. Priority of knowing: I know everything that I have learned.
2. Priority of learning: I have learned everything that I know.

The first axiom seems obvious. Learning would not really be learning if it did not lead to knowledge. Yet, this is not unproblematic. Learning has a temporal aspect. It involves a transformation from not knowing to knowing. If we simply learn a finite number of facts this is straight forward. If somebody tells me that Amsterdam is the capital of the Netherlands and I did not know that, then I have learned something. Of course I trust my source of information to speak the truth. He must be a trustworthy teacher. Even if that is the case things get more complicated if I try to learn an infinite number of facts in a finite time. Since Hume, philosophers know that this is logically impossible. One can never learn a general law on the basis of a finite number of observations. Even if I have seen millions of white swans, this does not allow me to draw the conclusion that the statement "All swans are white" is true. I only need to observe one black swan and my general law can be scrapped (Popper [1952]). The conclusion seems clear. Logically it is impossible to learn a infinite set on the basis of a finite number of observations. To put it in other words: we can learn facts, but we can not learn general laws. This would mean the end of science. Philosophers that endorse the first axiom implicitly wipe the problem of learning under the carpet: learning actually is remembering what you already know (Plato), you can only learn if knowledge is innate (Descartes, Chomsky), mathematical research is the discovery of what is already there (Hilbert, Gödel). Under axiom 1) scientific knowledge is only possible if one has what I call a participation theory of truth. The amount of knowledge of the human subject grows in time, but not by means of learning. The human mind seems to participate in the realm of truth and this participation allows us to separate true from untrue insights. It is clear that this theory of learning is less satisfactory.

So let's have a look at axiom 2) the priority of learning. From this perspective we seem to loose our grip on the concept of knowledge. Results that we have learned are preliminary, they can change, they have a statistical nature. In most cases learning leads to a hypothesis that only has a certain degree of plausibility. It does not seem to be a good idea to accept the derivation "The hypothesis that P is very probable, therefore I know that P" as valid. Knowing seems to be an absolute concept. The situation in which I testify in court that I know that John has killed Mary is very different from the situation in which I testify that it is very probably that John is the killer. Nevertheless we are willing to sentence somebody, even if we are not completely sure that he is guilty. Beyond reasonable doubt is a phrase that finds its philosophical roots in the work of Hume, who has chosen the second axiom as his starting point. This position leads to what I call a construction theory of truth. A supporter of this theory has two options. Either he admits that knowledge is a statistical phenomenon or he limits himself to knowledge that can be constructed out of elementary observations. This last option leaves very little room for science. Yet this position has been defended vigorously in the philosophy of mathematics by Brouwer and the early Wittgenstein. Traces of the first solution can be found in the works of Aristotle, Euclid, Locke, Hume and the members of the Wiener Kreis.

This short analysis shows that one could rewrite the history of philosophy with learning as a central theme. For a long time such a history would not contain much more than what I summarized above. Both axioms lead to unfortunate conclusions. A good choice is not really possible: a real philosophical problem. In the second half of the 20th century theoretical ideas developed rapidly mainly as a result of the application of insights from mathematical model theory and thermodynamics to an analysis of the phenomenon of learning.

3 The riddle of human efficiency

The mathematics of learning starts with the conception of learning as a game that is played between a student and a teacher. The game theoretical model of learning was first introduced by Gold in Information and Control in 1967. The problem that Gold studies is learning a language. The form of the game is as follows:

1. There is background knowledge. The teacher and the student agree beforehand on a(n) (infinite) class of possible languages, one of which is to be learned.
2. The teacher chooses one language from this class that he is going to teach.
3. A move of the teacher consists of the presentation of an example sentence from the language he has chosen. The teacher must be faithful. He is obliged to produce all possible sentences of the language in the limit at least once.
4. A move of the pupil consists of a guess of the language (a hypothesis) that the teacher has selected.
5. The game continues indefinitely. The pupil learns the language (wins the game) if he does not need to update his hypothesis from a certain moment on.

We can suggest the following practical interpretations of this abstract model:

- Theory of knowledge: the student is any human being, experience is the teacher, the class of languages is the set of possible theories about the world.
- Cognition: the student is the brain, the teacher is perception, the class of languages is the number of concepts that the human brain can learn.
- Methodology of science: the student is the scientist, the teacher is nature, the class of languages is the set of possible laws of nature.

For our purpose the abstract model is rich enough. The surprise of Golds paper was that he could prove that under these conditions, even if the game could

go on for ever, the student could not learn classes of languages of any interest with absolute certainty. This holds a fortiori for all natural languages that we all learn as children without much difficulty. Here we find an interesting problem that has not been solved adequately until this day and really only has become more urgent. One could baptize this problem the riddle of human efficiency. All our formal models of learning tasks indicate that learning, from a formal point of view, is next to impossible or at least extremely hard. The central issue here is that learning in Gold's model is distribution free, i.e. the only constraint is that every sentence of the language has a probability bigger than zero to be produced by the teacher. This allows for highly non-standard distributions on which one cannot expect general learning algorithms to converge.

In the last 40 years we have seen an overwhelming amount of amendments and adaptations of Golds model and theory construction certainly is not finished (See e.g. Angluin [1988]). The research concentrates on a number of issues: a restriction on the class of languages, using statistical techniques to select the hypothesis, richer interaction between the student and the teacher and the attitude of the teacher. In the original model of Gold the teacher only has to be reliable. He gives all the examples in a random sequence. It is easy to imagine that the teacher helps the student a bit, for instance by selecting simple examples first or by adapting the information content of the examples to the progress of the student. In this case we have a cooperative teacher. In its simplest form the cooperative teacher is nothing but a probability distribution over the set of examples that gives a higher probability to simpler examples. A student that studies under the guidance of a cooperative teacher has a much higher chance of selecting the right hypothesis with the help of statistical reasoning. Here we distinguish the contours of an interesting solution to the riddle of human efficiency in learning. Our efficiency might not be a achievement of human intelligence but more a reflection of the structure of the world in which we live. Nature around us is not completely random, it is organized and works as a cooperative teacher. Before we explore this concept further we need to develop a formal framework to study these concepts.

Learning as data compression

Suppose you switch on your television set and there are three different channels from which you can choose: random noise, a picture of a forest and a test image. From a computational point of view we can analyze these three data sets in the following way:

1. **Random noise**: this data set has a high complexity and therefore contains from a theoretical view a lot of information. Because the data set is the result of a random process it cannot be compressed in to a shorter description. This means that it does not contain any meaningful information. Nothing can be learned from it. These data sets are typical for systems that are in thermal equilibrium and thus have maximal entropy.

2. **The picture of a forest**: this data set has high complexity, but it also contains structure (the forms of the branches, leaves and trees repeat themselves). Therefore the image can be compressed in to a shorter description. We can extract meaningful information from the picture (e.g. the fact that we can distinguish 10 trees in the picture). We can learn a lot from this data set. These data set are typical for self-organizing systems that extract energy from the environment to create some form of order, e.g. living things, computational processes.

3. **The test image**: this data set looks very simple with regular geometrical shapes. It can easily be compressed and thus contains little information at all. Nothing much can be learned from it.

From these examples it is clear that we can learn the most interesting things from data sets that show a mix of structure and random elements. This is exactly the sort of data sets that one would expect in a computationally cooperative universe. Modern learning theory focuses on the analysis of this kind of data sets. The ambition is to find an optimal short description of the data set in terms of two new data sets:

- A structural part that described the regularities in the data set.
- An ad hoc part that describes the random elements of the data set.

Such a description is technically adequate if the length of the new description in terms of two data sets is (much) shorter than that of the original data set. In the literature this principle is known as the Minimum Description Length principle (Rissanen [1999]) or also as two part code optimization (Vereshchagin and Vitányi [2004]). Suppose that the picture of the forest has a size of 1280 x 800 pixels of 256 colors, than the uncompressed file will have a size of about 31 Mb. This is the amount of bytes we need to send via a communication channel if we want to communicate the contents of the file. As soon as we have an analysis of the meaningful content of the picture at our disposal we can summarize the content. In this way we get a sequence of interpretations of the picture in which more and more of the content is revealed:

Ad Hoc	**Structural**
A forest	A general description of forests
A set of 10 trees	A general description of the structure of a tree
A set of 3 birches, 4 willows and 3 oaks	A description of the specific structure of birches, willows and oaks
Etc.	Etc.

An important part of the research in learning theory concentrates itself on the development of algorithms that can separate a data set in an ad hoc and a

structural part. Many scientific problems can be reformulated in terms of a two part code optimization problem. I give a number of examples:

Data Set	Ad Hoc	Structural
Description of our solar system	Trajectories and size of the planets	Kepplers laws
Reuters Database	Structure and sequence of the individual sentences	English grammar
A composition by Bach	Structure and sequence of themes	Specifics of Bachs style
Human DNA	Structure and sequence of regions that code genes	A description of genes

Finding such a two part code optimization is usually not an easy task. One can formally prove that there is no universal learning algorithm for such a task. For some data sets we have good algorithms, for others not (yet). It is possible with a learning technique called genetic programming to derive the laws of Keppler from the observations of Tycho Brahe, but a good algorithm for learning a grammar on the basis of a corpus is not yet available (Adriaans and van Zaanen [2004]). In the following paragraphs we will develop a deeper understanding of learning as compression.

4 Learning, Computation, Information and Entropy

In this section we will develop a formal framework that helps us to understand learning better. The crucial step is the definition of the concept of information as something that could be objectively quantified. Prima facie it is immediately clear that the concepts of information and learning are related. If somebody tells me that Amsterdam is the capital of Holland and I did not know this, then I am getting new information and I have learned something. It seems impossible to learn without getting information and impossible to get information without learning. A discussion of the technical issues concerning the concept of information is not possible without an understanding of the concept of a Turing machine. In the next paragraphs we will first describe this basic notion and then turn our attention to the definition of information.

The Turing machine

In its simplest form a Turing machine is a device with a read-write head, a infinite working tape on which symbols can be read and written and a finite deterministic program for the manipulation of symbols. The only symbols needed are '1', '0' and 'b' (blank). The machine starts its calculation by reading input from the tape, its stops when a certain predefined final state is reached. Not all programs will stop. In fact Turing proved that there does not exist a program

that decides in all cases whether a certain machine will stop given a certain input (undecidability). The combination of machines and programs that stop in finite time is known as the *Halting Set*. This set could be seen as a transcendent object in computer science: we know it exists, but it can not be constructed. There are a number of reasons why Turing's device can claim to be associated with a universal scientific language. First of all the set of all possible programs for a Turing machine is the set of all possible binary strings $\{0,1\}^*$, which is equivalent to the set of natural numbers. Secondly, one can define a 'universal' Turing machine, that emulates all possible computations of all possible Turing machines by first reading a definition of a machine from the tape followed by the definition of the program and the execution of the program on the emulated machine. This allows us to interpret the Turing machine as a universal computing device. Thirdly, all the current definitions of the concept of computation (Lambda calculus, combinatorial logic, recursive functions, etc.) are known to be Turing equivalent, i.e. can be emulated on a Turing machine. This fact has lead to the formulation of the so-called Church-Turing thesis, which states everything computable is computable on a Turing machine. It is hard to imagine how this claim could ever be verified. In the worst case it is destined to be an unproven metaphysical claim for ever. The thesis could easily be falsified by a conception of calculation that can not be emulated on a Turing machine, but sofar these conceptions of computation escape our imagination. From a transcendental point of view the Turing machine encapsulates fundamental notions: *The local physical storage and processing of a finite set of discrete symbols as a sequential finite discrete process in time according to a finite set of (deterministic) rules.* The apparent universality of these notions lead to what one might call the central working hypothesis of modern computer science:

Conjecture 4.1 *Any finite discrete system or process can be described in terms of a program for a Turing machine.*

Personally I expect this claim to be falsified (or at least amended) somewhere in the future, but for the moment it gives the foundation for a methodological research program that is rich in perspectives and far from exhausted. It defines a universal scientific methodology. For any system X we have to ask ourselves the fundamental question: is X a finite discrete system? If so we can apply our methodology and try to construct an adequate program to model it. The decision to consider a certain phenomenon X (say a financial administration, turbulence around a sail, human consciousness, the human cell, a black hole or the universe as a whole) to be a finite discrete system can be controversial from a philosophical point of view and require a separate philosophical motivation. These questions are not part of our current analysis. For the moment I aim at clarification of the central concepts and not at an analysis of their applicability.

The association with the old philosophical ambition of a mathesis universalis is immediately clear from the Turing equivalence of recursive functions, which lead to the following collorary:

Corollary 4.2 *Any finite discrete system or process can be described in terms*

of operations on natural numbers. [7]

This analysis of Turing machines does not lead to a theory of information. It is a theory neutral conception of manipulation of binary strings. In order to determine what kind of information, and how much of it, is contained in these strings we need separate definitions. Even within this context there are a number of competing conceptualizations of the notions of information that need to be treated here.

Shannon Information and optimal codes

The idea that the frequency of a letter is associated with the information it contains (or its value) is well known to any person who solves a crossword puzzle or plays Scrabble. If one knows that a word contains a 'z' this is more informative than an 'e' because there are less words with a 'z'. This 'information' about the 'z' implies a bigger reduction of the search space. The crucial insight that has lead to a mathematical theory of information is formulated by Shannon (Weaver and Shannon [1949]). Here the information content of a message is defined in terms of its probability:

Definition 4.3 *The Shannon information contained in a message x is $I(x) = \log 1/P(x) = -\log P(x)$,*

where $I(x)$ is the amount of bits of information contained in x and $P(x)$ is a probability distribution ($0 \leq P(x) \leq 1$). Note that[8]: If $P(x) = 1$ then $I(x) = 0$. $I(x \text{ and } y) = I(x) + I(y)$.

From a philosophical point of view it is important to note that Shannon information says nothing about the meaning of the messages, nor about their epistemological status. If x is a message and $P(x) = 2^{-3}$ then the amount of information contained in x is three bits and an optimal code for x would use three bits, say 001. Apart from this x could have any meaning, varying from "John has passed his exam" to "Goldbach's conjecture is true". In itself this is strange. We are inclined to say that if we get the information that John passed his exam from a reliable source we consequently *know* that John passed his exam. A simple bit code like 001 does not convey this information. Apparently there are meanings of the term 'information' that are not fully covered by Shannon's definitions. Shannon himself, by the way, would be the first to acknowledge this. Also there is no straightforward translation of Shannon's definitions in to a theory of knowledge. A valuable attempt to fill this gap is made by Dretske. Dretske [1981] The least one can say is that, on top of the formal definitions that are offered by Shannon, the factual information that is transferred from a sender to a receiver is dependent on the context of the dialogue and on the background knowledge shared by parties involved in the exchange of messages.

[7] Wolfram states a related notion that he calls the Principle of Computational Equivalence: "...whenever one sees behavior that is not obviously simple ... it can be thought of as computation of equivalent sophistication" (Wolfram [2001], p. 5).

[8] *log* is used for log_2

A second observation that is philosophically relevant is that Shannon information as such is independent of the notion of a Turing machine. Shannon defines information in terms of bits and Turing machines operate on strings of zeros and ones that could be interpreted as bit strings. In these terms Turing machines could be seen as information processing devices, but this is only a very weak connection. Shannon's notion of information and Turing definition of computation seem to orthogonal. Shannon uses the notion of a bit to measure amounts of information, but his theory does not say anything about the amount of information that is stored in a string of bits itself.

The concept of Shannon information only makes sense in the context of a set of potential messages that are sent between a sender and a receiver and a probability distribution over this set. If we have such a setting we can design an optimal code system. Suppose X is a set of messages $x_i(I = 1, \ldots n)$ the **communication entropy** of X is: [9]

$$H(X) = -\sum_{i=1,n} P(x_i) \log P(x_i)$$

. The **Maximal entropy** of a set of n messages, if $P(x_i) = 1/n$ for each I:

$$H_{max}(X) = -n(1/n)\ log\ (1/n) = \log n$$

. The **Relative entropy**: $H_r = H/H_{max}$, the **Redundancy**: $1 - H_r$, the **Optimal code** (that minimizes the expected message length) assigns $-logP(x_i)$ bits to encode message x_i. One finds an extensive discussion of these definitions in the chapter by Harremoës and Topsøe. The notion of optimality of a code system is associated with the idea of compression of a set of messages. Suppose, for the sake of argument, that we want to develop an optimal code for a certain book, say Dickens' "A Tale of Two Cities", and that we simplify the task to finding an optimal code for an alphabet of 26 letters. [10] We can code each of the 26 letters with a standard length of 5 bits. A set of messages in which the frequency of each letter would be equal (e.g. 1/26) has maximal entropy. Of course such a set would contain only nonsense. It could not be normal English since the frequency of letters in English varies greatly. Therefore a standard 5 bit code is redundant and can be optimized. We can assign shorter codes to more frequent letters. Giving up the fixed code length implies that our code has to be *prefix free*: no code can be a prefix of any other code. Standard Huffman code provides an optimal solution for this problem. Using Huffman code one can compress "A Tale of Two Cities" 0.81 bit per character comparison with the 5 bit code. We can ask ourselves if Huffman code is the best solution for compressing a book. In a sense it is, if one sticks to compression of characters, but there is no reason to do this. One could try to compress words instead or maybe one could use an analysis of idiosyncrasies of Dickens' style. This poses an interesting theoretical problem: what would be the theoretical shortest code

[9] This definition is exactly equal to the definition of Gibbs entropy in thermodynamics. See the chapter by Bais and Farmer in this book.

[10] This example is discussed extensively by Harremoës and Topsøe.

for "A Tale of Two Cities"? In order to find an answer for this question we have to turn our attention to a different definition of the concept of information that is intricately related to the notion of a Turing machine: Algorithmic Information.

Algorithmic information

We have seen that with the theory developed by Turing we can define a universal Turing machine. In fact there is an infinite number of such universal Turing machines, so let us select a standard (small) one and call it U. The input of U consists of two parts: a definition of a special Turing machine T_i in prefix code, followed by the input code, or data D for T_i. Observe that, using Huffman code, we can create a program that reproduces "A Tale of Two Cities" as output on U. The crucial insight is that it is easy to construct a Turing machine that decodes Huffman code. Let $D_{ToTC,Huf}$ be the Huffman code for "A Tale of Two Cities" and let T_{Huf} be a Turing machine that decodes Huffman code in the standard prefix free input format of U. The text of "A Tale of Two Cities" can be coded as

$$U(T_{Huf} + D_{ToTC,Huf})$$

When confronted with the input $T_{Huf} + D_{ToTC,Huf}$ our universal machine U will first read the definition of T_{Huf}, reconfigure itself as an interpreter for Huffman code and then start to interpret $D_{ToTC,Huf}$ resulting in the text of "A Tale of Two Cities" as output. The bit string $T_{Huf} + D_{ToTC,Huf}$ can be seen as a program for the text of "A Tale of Two Cities". Let $|D|$ be the length in bits of the data set D and let $D_{ToTC,5bit}$ be the 5 bit code for "A Tale of Two Cities. We will have:

$$|T_{Huf} + D_{ToTC,Huf}| < |D_{ToTC,5bit}|$$

Given the fact that a Turing machine for interpreting Huffman code is not complicated, the set $T_{Huf} + D_{ToTC,Huf}$ will be shorter than the original 5 bit code for "A Tale of Two Cities". In this way we have created a computer program that generates the text of "A Tale of Two Cities" on a universal Turing machine. The bit code of this program is shorter than the original text. We could go on and try to find more clever code systems that compress the text even more. Such a code system, say $T_{CodeSystem_i}$ could make use of the frequency of words in the text, knowledge about the grammar of English and idiosyncrasies in the style of the author. Such a code system would be 'better' than the Huffman code if:

$$|T_{CodeSystem_i} + D_{ToTC:i}| < |T_{Huf} + D_{ToTC,Huf}|$$

where $D_{ToTC:i}$ is the text encoded in the new code.

We can now answer the theoretical challenge from the previous paragraph: the theoretical shortest code for "A Tale of Two Cities" would be the shortest program that generates this text on U. In order to find this program ideally, what we have to do is enumerate all possible programs for U, test them, and select the shortest that generates "A Tale of Two Cities". Alas this is impossible

because of the uncomputability of the halting set. We know that such a program exists, but it remains an intensional object.

This fact gives rise to a different definition of the concept of information (Li and Vitányi [1997]). The descriptive complexity of a string x relative to a Turing machine T and a binary string y is defined as the shortest program that gives output x on input y:

$$K_T(x|y) = \min\{|p| : p \in \{0,1\}^*, T(p,y) = x\}$$

One can prove that there is a universal Turing machine U, such that for each Turing machine T there is a constant c_T, such that for all x and y, we have $K_U(x|y) \leq K_T(x|y) + c_T$. [11] This definition is invariant up to a constant with respect to different universal Turing machines. Hence we fix a reference universal Turing machine U, and drop the subscript U by setting $K(x|y) = K_U(x|y)$. We define:

Definition 4.4 *The Prefix Kolmogorov complexity of a binary string x is $K(x) = K(x|\epsilon)$. That is the shortest prefix free program that produces x on an empty input string.*

Kolmogorov complexity is a competing notion of information. It allows us to assign a complexity to individual strings and data sets.

A unified view on Shannon information and Kolmogorov complexity

We are now in a position to evaluate the difference between Shannon information and Algorithmic information, i.e. Kolmogorov complexity. Suppose we have a data set encoded in bits, say a five bit code of the text of "A Tale of Two Cities". We can analyze this set from two perspectives:

- From a Shannon perspective as a *collection of messages*. In this we can construct an optimal code using variation in frequency of the messages. This leads to a relative compression of the set of messages that can be computed. More frequent messages get shorter codes and contain less information. We could call this concept of information *relative*.
- From a Kolmogorov perspective as a *single message*. In this case relative frequency has no meaning, but there exists an optimal compression of the message in terms of the shortest program on a Turing machine. The length of this program is an absolute measure for the amount of information contained in the message. This program is an intensional object and can not be computed as such. Messages that are highly compressible contain little information. This could be seen as a concept of *absolute* information.

As an example, suppose we have a bit string 01010101010101010101010101. We can *recode* this string in Shannon's sense as '01'=1;11111111111111, or

[11] For an extensive discussion of these definitions, see the chapter by Grünwald and Vitányi in this book.

we can *reprogram* it in Kolmogorov's sense as `for x = 1 to 13 write '01'`. Both structures are shorter than the original code reflecting the fact that the string shows a regular pattern. In this case both the Shannon and the Kolmogorov compression do their work. In my view both algorithmic information and Shannon information are different mathematical guises of one and the same concept of information that is associated with entropy of data sets.

Claim 4.5 *Information is associated with the entropy of data sets. Data sets with low entropy can be compressed and contain less information than data sets with maximal entropy, which cannot be compressed and contain exactly themselves as information. There are various ways to explain these relations mathematically.*

Shannon information starts with a segmentation of the set. In the limiting case where we have very few segments, or only one, Shannon's theory collapses in to Kolmogorov's conception of information. Kolmogorov's conception of information is more powerful, but the price we have to pay is threefold: it is non-constructive, therefore it can only be approximated and it is asymptotic.

Lemma 4.6 *The concepts of Kolmogorov complexity and Shannon information are equivalent in terms of predicting incompressibility of data sets with maximal entropy.*

Proof: In Shannon's conception a set of messages can not be compressed if they all have equal probability. Suppose we have a sequence of k messages with maximal entropy based on a code system of 2^n code words of n bits, then this is equivalent to a random string of $l = kn$ bits and thus it can not be compressed in Kolmogorov's sense. Suppose, conversely, that we have a random bit string $l = kn$ bits with l fixed, then for each segmentation of l in k messages the entropy is maximal thus it can not be compressed in Shannon's sense.

Note that the difference between Shannon information and Kolmogorv information can be seen as a difference in graining. Kolmogorov complexity is coarse grained giving the whole set of messages a complexity in one shot. Shannon information is fine grained, it calculates the information for individual messages first and then establishes an entropy for the whole set. Given the equivalence of Shannon information and Kolmogorov complexity one would expect that also in the limiting case of considering a bit string as one unsegmented message it is possible to assign a probability to it. This is indeed the case. In Shannon's case we reason from probabilities to entropies, in the Kolmogorov world we derive probabilities from entropies. Using results of Solomonoff(Solomonoff [1997, 2003]) and Levin we can define an a priori probability of a finite binary string.

Definition 4.7 (Solomonoff, Levin) *The universal a priori probability $P_U(x)$ of a binary string x is*

$$P_U(x) = \sum_{U(p)=x} 2^{-|p|}$$

This is the sum of the probabilities of all the programs that generate x on a universal Turing machine on an empty input string. Thus strings with a low Kolmogorov complexity, i.e. the ones that are compressible, get a higher a priori probability. Associated with a universal a priori probability we expect to get a universal distribution. We can define a semi-measure along these lines. A recursively enumerable semi-measure μ on N is called universal if it multiplicatively dominates every other enumerable semi-measure μ' i.e. $\mu(x) \geq c\mu'(x)$ for a fixed positive constant c independent of x. Levin proved that such a universal enumerable semi-measure exists. Since there might be more we fix a universal semi-measure $\mathbf{m}$(x). The semi-measure $\mathbf{m}$(x) converges to 0 slower than any positive recursive function which converges to 0. Of course $\mathbf{m}$(x) itself is not recursive. We now give without proof a theorem that relates all these concepts with each other:

Theorem 4.8 (Levin)

$$-\log \mathbf{m}(x) = -\log P_U(x) + O(1) = K(x) + O(1)$$

The universal distribution has quite wonderful qualities and its philosophical relevance has hardly been explored up till now.

Thermodynamics, Information and Computation

It is clear that the study of information and computation is related to concepts of thermodynamics on a fundamental level. The first law of thermodynamics states that energy in a closed system is conserved. The second law states that the entropy of a closed system can never decrease. After a certain time a closed system will reach an equilibrium in which the entropy is maximal. Another way of phrasing the second law is that self-organization is not possible without external energy.

As the entropy of a set of messages grows, so does the set of accessible states and so does the number of bits that we need to identify those states (according to Boltzmann the formula entropy was simply $S = \ln w$, where w is the number of accessible states, this is equal to the maximum entropy in Shannon's definition). Consequently in a closed system, when the entropy grows, the amount of information stored in the system grows. A closed system can increase its internal information without exchange of heat with the environment. This is actually what is happening in our universe. There is much more information in the universe now, than there was at the moment of the big bang (otherwise it would be a dull place). At the same time the universe is getting more and more improbable.

A thought experiment can help here. Think of a bit string as a gas in a one dimensional container (say 0's are spaces and 1's molecules). If the bits are allowed to move freely through the space starting from any configuration they will eventually reach an equilibrium state in which the Kolmogorov complexity of the accessible states is maximal. These states are exactly the ones in which the bits contain maximal information (in terms of Kolmogorov complexity).

Random bit strings contain the most information, have the highest entropy and correspond to a thermal equilibrium. [12]

All this is quite counter intuitive. If we dissolve milk in coffee, or we spoil sugar in sand we feel we loose possibilities. It seems strange to assume that noise on a channel is actually the richest source of information possible. The reason for our unease seems to be the fact that high entropy is the normal situation in the universe. Order (i.e. low entropy) is more interesting since is needs to have a specific cause. High entropy does not point at specific causal processes of any interest. Low entropy is a sign that somebody or something redirected energy to a system. That is the reason why, when we want to detect life in outer space, we scan the sky for signals with less then maximal entropy. It is therefore better to speak about meaningful information. In order to be meaningful to us, a set of messages has to have some structure and consequently have less than maximal entropy. This concept of meaningful information in a system is from a thermodynamical point of view related to the free energy in the system and from a learning view to minimum description length and two part code optimization.

Thermodynamics therefore has interesting consequences for the physics of computing. A universe in which can be calculated has to obey the following conditions:

- It must be stable enough to **store information**. Structures should have a certain stability; identity over a certain period of time should be guaranteed. This points at relatively low entropy. In a system in thermodynamic equilibrium structures would not be robust enough to store information over a certain period of time.

- There must be enough free energy to **process information**. There must be reversible processes that facilitate the transition between stable states: i.e. there must be mechanisms to flip bits. Because of the second law of thermodynamics such a system will, according to the Landauer principle[13], always require energy to erase information. This is quite subtle. Erasing of information requires energy, creation not necessarily. This condition implies more than minimal entropy. Computation can not exist in systems with extremely low entropy, e.g. computation at zero degrees Kelvin is not possible.

Computation seems to presuppose some kind of state of intermediate non equilibrium state of entropy.[14] Luckily we live in a universe that satisfies these conditions exactly. This is no surprise, because in a universe that does not offer these possibilities intelligent life would not be possible. This is a variant of the antropic principle (Hawking [1988]). The hypothesis of the cooperative however goes deeper because she states that such a universe would be easy to learn. It

[12] It is possible to develop a thermodynamics of bit strings along these lines.

[13] See the chapter of Bais and Farmer in this book.

[14] This goes against the interpretation of Lloyd and Ng (Lloyd and Ng [2004]) who consider almost any physical process as a computer, e.g. black holes and pure plasma. In these cases it is better to speak of computational processes than of computers.

is a number of random processes, but these processes are necessarily of limited complexity.

Out of these observations the following picture emerges. A deterministic computer is simply a Laplacian system that in itself cannot add information to the universe. Its future is completely determined by its initial conditions. Still a deterministic computer can easily use energy to erase information and thereby reduce the amount of information in the subsystem (say its tape). The total entropy in the universe will still grow as a result of this action. For a subjective observer however the situation is different. He might not know whether a certain computation will finish. If he observes that the computational process comes to a halt this certainly adds to his information, even if he lives in a Laplacian universe.

Suppose on the other hand that a statistical observer can only make measurements of a certain granularity. He can for instance measure the local density of bits on the tape with a certain accuracy, but not observe individual bits. In such a case the subjective entropy generated by a deterministic computing process can be much bigger than the entropy of the initial conditions. Suppose that the computer writes the binary expansion of the number e on the tape. This is a data set with very low entropy, but for such a statistical observer it cannot be distinguished form random noise (since he cannot identify the individual bits). Here we seem to cross the border from theory of computation to thermodynamics. Very much the same thing happens if we see the generation of a fractal. This is a data set of very low entropy, but to our subjective eye full of interesting details. A non-deterministic computer adds information to the universe with each randomized computing step it takes.

As a last note observe that thermodynamics only works for systems in a state equilibrium. Computing systems tend to specifically not in an equilibrium so the applicability of classical thermodynamics for the understanding of computing processes is limited. At the moment we are missing a theory that helps us to understand these matters adequately. The following theoretical observations give an initial outline of such a theory.

A universal a priori near optimal Shannon code based on Kolmogorov complexity

Levin's theorem allows us to explore the relation between Shannon information and Kolmogorov complexity at a more fundamental level. We define the standard bijection b between the set of binary strings $\{0,1\}^*$ and the set of natural numbers N as

$$b(0,\epsilon), b(1,0), b(2,1), b(3,00), b(4,01), \ldots$$

Where ϵ denotes the empty word. We can define the function $S : \{0,1\}^* \to \{0,1\}^*$ as:

Definition 4.9 $S(x) = \min_{i \in N}\{p : b(i,p), U(p,\epsilon) = x\}$

Here U is a universal Turing machine. S associates each binary object x with the first program that produces x on U with empty input.

Corollary 4.10 *S is a universal a priori near optimal code associated with* $\mathbf{m}$ *for binary strings in Shannon's sense.*

Proof: According to Shannon an optimal code for x given $\mathbf{m}$ would be $-\log \mathbf{m}(x)$ bits long. According to Levin we have $-\log \mathbf{m}(x) = K(x) + O(1)$. But then $S(x)$ is such an optimal Shannon code, because by definition $|S(x)| = K(x)$ since $S(x)$ is the first, and thus the shortest, program that produces x on U. The code is near optimal, because of the factor $O(1)$ in Levin's theorem. $S(x)$ will always be maximally $O(1)$ removed from the factual optimal code.

The function S is interesting because it brings the concepts of Shannon information and Kolmogorov complexity together. On one hand $|S(x)|$ is the Kolmogorov complexity of x, on the other $S(x)$ is an optimal a priori code for x. Of course S can never be computed, but suppose that some Platonic oracle would give us S. In that case we would have a universal a priori solution to the problem of induction. $S(x)$ reflects *any regularity (e.g. deviation from maximal entropy, i.e. compressibility) that can be expressed solely in terms of the internal structure* x. Observe that $S(x)$ will itself always be random (and thus incompressible) because it is the *first* program that computes x. If $S(x)$ would be compressible, it would itself have been identified much earlier by S. It is important to note that, although S can not be constructed, it nevertheless really exists. S is the closest we can get to a universal language of science, given the current state of research in computer science.

To give some examples. S would make it easy to find binary expansions of transcendent numbers like π and e. There are simple programs for these extensions. In fact S would identify almost *any* discrete object of *any* mathematical interest for us. On top of that S would give us an optimal code for the text of "A Tale of Two Cities" and indeed of any other conceivable poem, novel, piece of music, movie or any work of art in digital code. The same would hold for any digital data set that scientific inquiry could produce. S would 'explain' the regularities and idiosyncrasies of these data sets in so far as they can be expressed in terms of deviation of maximal entropy.

Intensive and extensive data sets

A very interesting consequence of having S would be that we are capable of measuring the scale invariance of complexities and entropies. A little thought experiment will help. Suppose that we study some segment L of length l, starting at the p-th bit, of the binary expansion of a transcendental number, say π. Since we are studying an expansion of π the Kolmogorov complexity of the sequence is low. In the sense of lemma 4.6 we could analyze this as a sequence of $l = kn$ bits, i.e. k messages based on a code system of 2^n code words of n bits. The total measured complexity of L using S with granularity n could be defined as:

$$K(L)_{S,n} = \sum_{i=0}^{k-1} S(x_{(i\times n)+1}, x_{(i\times n)+2}, \ldots x_{(i\times n)+(n-1)})$$

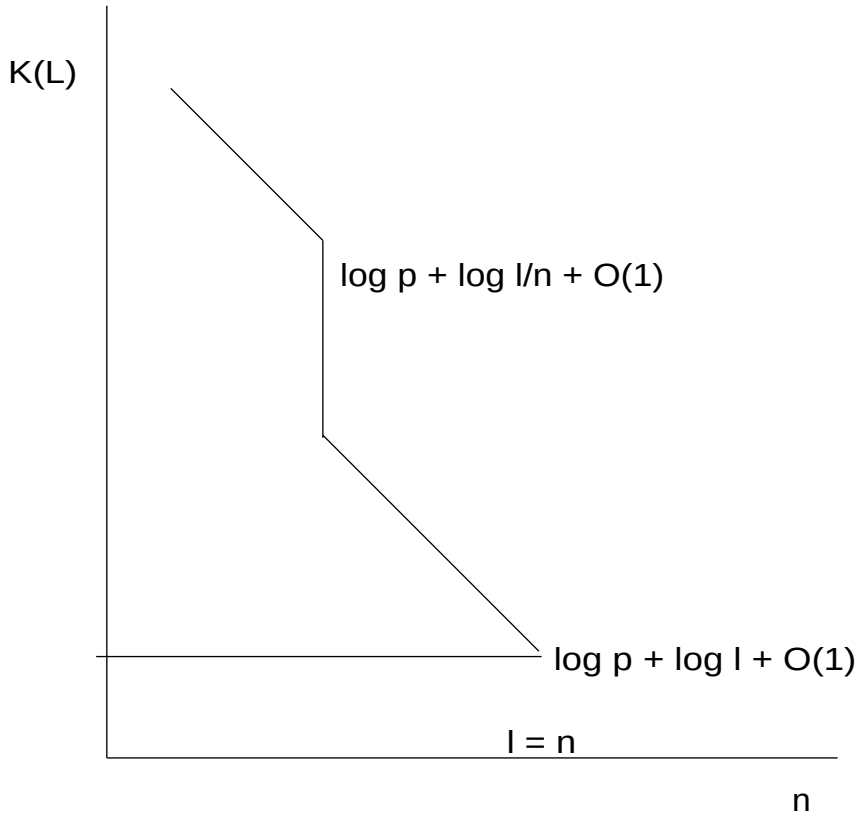

Figure 1: The size of $K(L)_{S,n}$ in relation to the granularity n while sampling a segment of π

If we plot the size of $K(L)_{S,n}$ in terms of the size of n we will see the following effect: for small n the function $K(L)_{S,n}$ will show a slow decrease that will be linear in n. This is because of the diminished overhead of S per segment. For small n all segments will be random for S, because of the transcendentality of π. At a certain point, 'close' to $\log p + \log l/n + O(1)$, the value of $K(L)_{S,n}$ will drop suddenly. [15] This is exactly the point where n is big enough so that S starts to 'sense' the compressibility of L. For $n = l$ the function $K(L)_{S,n}$ will land at the value $\log p + \log l + O(1)$. What this amounts to is that for certain data sets, e.g. bit representations of transcendental numbers (but there are many others), complexity (and consequently entropy) is *non-extensive*. Another way of putting this is that the Shannon entropy of the collection of messages diverges from the Kolmogorov complexity as a measure of entropy for the set as a whole. Local

[15] The $\log p$ gives us an index in L, $\log l/n$ code the length of the individual segment and the $O(1)$ term contains the program for π. This information is sufficient to describe any substring in L.

estimates of the complexity do not tell us anything about global complexity and consequently complexities of various regions of the data set can not be added to get a global complexity estimate. The complexity of these data sets is not robust under statistical operations and under re-scaling of the code system. [16] Clearly for the application of efficient learning algorithms the non-extensive complexity of such data sets is an unsurmountable barrier. No algorithm can compress data sets that look random from the outside but are in fact highly compressible.

Uncompressibility and extensiveness are in fact the same notions, as is clear from the following analysis. A data set D is extensive if the sum of the complexity of two arbitrary disjoint subsets A and B equals that complexity of the union of that set: $K(A) + K(B) = K(A \cup B) + O(1)$. This is only the case if D does not contain any redundancy i.e. if D is random. On the other hand, suppose that D is very compressible. If we know A already, then B would add no information, i.e. $K(A) + K(B) = K(A) + \log|B| + O(1)$. In other words B would only add its own size to our knowledge. This is for instance the case when D contains extremely simple regular patterns. This suggests the following definitions:

Definition 4.11 *A bit string D is **extensive** for a sample granularity g if for each substring $A \in D$ such that $|A| \geq g$ we have $K(A) > |A| - O(1)$. A bit string D is **intensive** if for each substring $A \in D$ such that $|A| \geq g$ we have $K(A) < \log|A| + \log|D| + O(1)$. **Sub-extensive** data strings have $|A| \gg K(A) + O(1)$ and **super-intensive** strings have $K(A) \gg \log|A| + \log|D| + O(1)$.*

Sub-extensive data sets are the ones from which we can learn something. The borderline between extensive, sub-extensive, super-intensive and intensive data sets is blurry, but the general idea stands. If we sample an extensive data set we really get value for money, every bit counts. But there is a price to pay. The information is completely random. Nothing can be learned from this set. This corresponds with the picture of random noise at the television set that was discussed earlier in this chapter. On the other end of the spectrum we find the picture of the test image: this data set is almost totally intensive. It is a simple repeatable pattern for which we need only the information about the number of repetitions to encode it. Extensiveness corresponds to maximal randomness, intensiveness to maximal redundancy. Figure 1 shows that we can make each string extensive by taking a small granularity. This corresponds to the fact that, even if a data set is very regular, there is a learning phase in which we have to analyze the pattern itself. At this time the data set cannot be distinguished from a random one. A finite program producing an infinite data set has to go through loops. If we cannot compress the data set on the basis of samples that are in the order of the complexity of a loop of the program that generates the data we are in trouble. Because the increase in information after this phase will be only logarithmic. So if we have not spotted the regularity after, say 10, loops

[16] The custom in thermodynamics to take the averages of values in the sample regions is just one specific form of recoding.

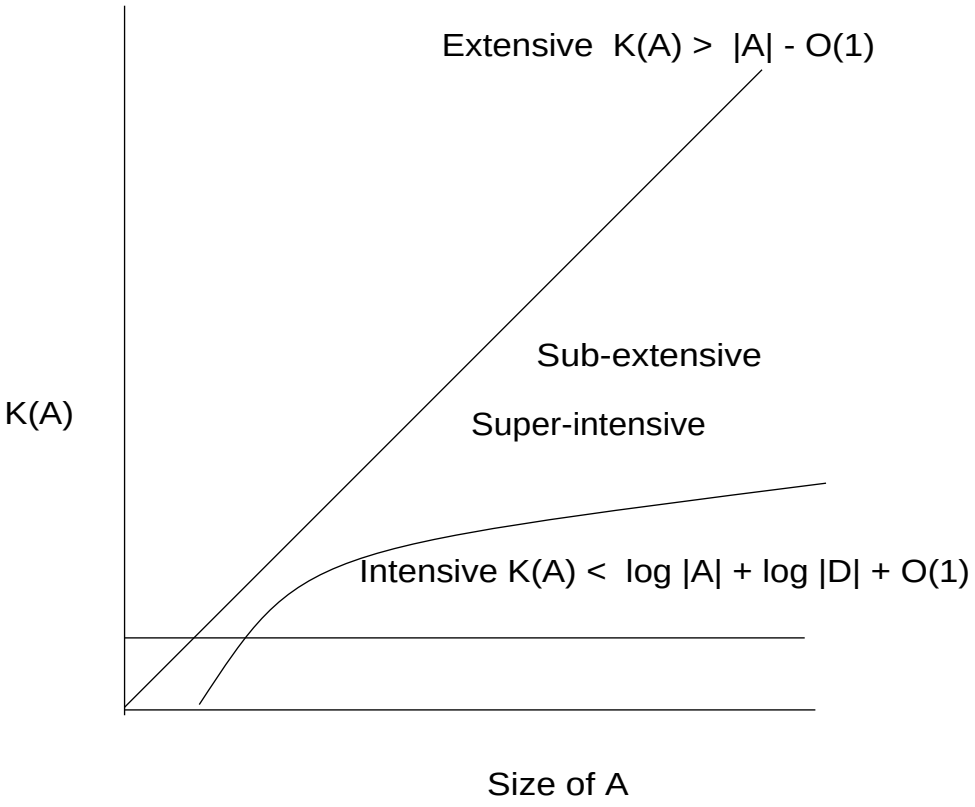

Figure 2: The relation between extensive, sub-extensive, super intensive and intensive strings

then we will probably never spot it because the only new information we get from x repetitions is of size $\log x$. This gives rise to the following claim:

Claim 4.12 *From the point of view of intelligent systems of a certain complexity, nature is by necessity shallow. Intensive data sets can either be learned by an intelligent system (a resource bounded learning algorithm) that is of the order of the complexity of the algorithm generating the data set, or not at all.*

From completely intensive strings we can learn only their generating program and their size. The program generating an intensive string can be seen as its **intension**.[17] Intensive data sets asymptotically have their size as their most defining characteristic. Extensive data sets do not have an intension, or to say this in other words: they only describe themselves. Their extension is their intension. Super-intensive data sets contain more information, but this

[17] Here we have a computational equivalent of Platos notion of an idea. The intension of an object is the program generating it.

might be just noise. They are non random, but not completely regular either. From a physical point of view they are associated with systems that are in a non equilibrium state. It is the kind of information that we find in the picture of the forest on our television screen. The trees are generated by a program and thus have regular specific features. But the program is not completely deterministic. Individual trees show random variation. It is interesting to characterize sciences in terms of the nature of their data sets. Data sets of mathematicians and physicists are close to intensive. Data sets of the humanities are super-intensive. The eternal question whether history repeats itself, can be answered by stating that history is sub-extensive and super-intensive. There are patterns but they will never repeat themselves exactly. In physics we have explanation and prediction exactly because the data sets are intensive.

A consequence of this analysis is that the amount of randomness we observe is dependent on the granularity of our measurements. In one sweeping statement one might say: randomness has a scale. Suppose we are looking at a movie of a hand flipping a coin. [18] At normal speed we are looking at a random (or at least a very complex) process. This data set certainly has extensive elements. Note that the data set itself in this case is not random. It is a movie of coin flipping that contains a lot of information. We could for instance learn a lot about Newtonian mechanics if we analyze it at an appropriate scale. Now suppose that we slow the movie down extremely, say we stretch out one second to a million years. In this case the movie will be rather dull on a human scale. It will be close to a intensive process that contains very little information. On the other hand if we speed the movie up so that a million years is compressed in to one second. Then again the movie would on a human scale be reduced to a meaningless grey blur that contains no information. On this scale the data set would again be intensive. The important thing to notice is that the data set contains the most information if we sample it at a granularity where the extensiveness is maximal. Both at a larger and at a smaller granularity we will lose information. In short: even randomness has a scale. Every form of randomness necessarily can be observed at a granularity in which it is in equilibrium. When we see smoke dissolve in the air, then on a human scale we observe increase of entropy, on a molecular scale the increase does not exist and on the scale of, say the solar system, the effect is too small to notice. An optimal analysis of a data set involves finding a granularity that optimizes the randomness of the data. [19]

[18] Suppose also that this hand does not belong to Persi Diaconis, the well known matematician/magician that has proved that coin flipping is actually a deterministic process. Some of the material in this paragraph is influenced by the lecture that professor Diaconis gave on the occasion of receiving the Van Wijngaarden award at CWI in 2006.

[19] This insight is related to Jaynes' maximal entropy principle and the minimal randomness deficiency principle to be discussed later. There is a further analogy with thermodynamics, where we find exactly the same scaling issues. Suppose that we have a number of gas particles in a isolated container at low entropy. After some time an equilibrium will be reached. On a micro scale the entropy can not have increased because the evolution of particles in the container is determined by simple deterministic Newtonian physics. Macroscopic measurements however will show an increase in entropy. Just like our example of the binary expansion of π,

Researchers in machine learning are familiar with the idea that certain phenomena can only be explained at certain scales. Some structures can only be learned when the data set is sampled with a certain granularity. [20] This can also be observed in the text of "A Tale of Two Cities". When we only sample individual bits of this data set no useful information emerges. When we sample letters, we can make good statistical estimates based on frequency. This is already somewhat harder for words and next to impossible for sentences, leave alone paragraphs or chapters. There is a certain granularity that reveals the structure of the text optimally.

A deeper analysis of these kind of phase transitions and their meaning for learning algorithms is necessary, but it is clear from this short analysis that the analogy between information and thermodynamics can be carried further than is commonly accepted.

Induction and Minimum Description Length

Let us have a closer look at the relation between S and the problem of induction. In one special guise induction amounts to selecting the most probable hypothesis to explain a given data set. In terms of Bayesian learning this task can be formulated as follows. Mitchell [1997] The **prior probability** of a hypothesis h is $P(h)$. Probability of the data D is $P(D)$. The **Posterior probability** of the hypothesis given the data is:

$$P(h|D) = \frac{P(h)P(D|h)}{P(D)}$$

Theorem 4.13 *Suppose that $h, D \in \{0,1\}^*$, i.e. both the data set and the hypothesis range over the full class of finite binary strings. Selecting the* **Maximum A Posteriori hypothesis (MAP)** *to explain D, amounts to selecting the hypothesis that minimizes the length in bits of*

$$S(h) + S(D|h)$$

.

Here $S(h)$ is the universal optimal Shannon code for the hypothesis and $S(D|h)$ is the universal optimal Shannon code for the data set given the hypothesis.

the data set will have low complexity at micro level and appear to be random at lager scales. In a strictly deterministic universe randomness takes the form of coarse grained undecidability.

[20] This was one of the more interesting results of the Robosail project, an attempt to use machine learning techniques to learn to sail automatically that I started in 1998 (van Aartrijk et al. [2002]). Measurements of almost all relevant human concepts like 'wave', 'gust of wind', 'change of wind direction' and 'wind strength' were dependent on selecting an adequate granularity for the measurements. What you subjectively experience as a wave is dependent on the size of your boat. Some of the conceptual distinctions used by sailors depend on sophisticated phase transitions in chaotic media that were only observable at certain scales. This holds for instance for the distinction between light air (laminar flow) and breeze (turbulent flow). In the final system we implemented learning agents that were living in a variety of time scales: 10 Hz, 1 Hz, 10^{-3} Hz, etc.

Proof:

$$h_{MAP} \equiv argmax_{h \in H}\ P(h|D)$$

$$= argmax_{h \in H}\ (P(h)P(D|h))/P(D)$$

(since D is constant)

$$= argmax_{h \in H}\ (P(h)P(D|h))$$

$$= argmax_{h \in H} \log P(h) + \log P(D|h)$$

$$= argmin_{h \in H} - \log P(h) - \log P(D|h)$$

(Since $h, D \in \{0,1\}^*$ and according to Shannon $-\log P(h)$ is the optimal code for the hypothesis and $-\log P(D|h)$ is the optimal code for the data given the hypothesis.)

$$= argmin_{h \in H} S(h) + S(D|h)$$

This result is closely related tot the so-called:

Definition 4.14 The Minimum Description Length principle (MDL): *The best theory to explain a set of data is the one which minimizes the sum of*

- *the length, in bits, of the description of the theory and*
- *the length, in bits, of the data when encoded with the help of the theory*

This principle was first formulated by Rissanen. Rissanen [1999] Research in this domain is far from finished and these concepts are still the object of fierce debate (Domingos [1998] Domingos [1999]). A common misconception is the idea that the minimum description length principle can be transformed in to a methodology for the construction of a sequence of improving theories by means of an incremental compression of the data set. Suppose that S_i, h_j, Sp and h_q are arbitrary coding schemes and hypotheses such that:

$$|S(h) + S(D|h)| < |S_i(h_j) + S_i(D|h_j)| < |S_p(h_q) + S_p(D|h_q)| < |D|$$

Although h is the best theory it is not necessarily the case that h_i is better than h_q. This could for instance be guaranteed if $S = S_i = S_p$, i.e. when the code is optimal (Adriaans and Vitányi [2005]).

Translating these observations to the domain of methodology of science gives us a number of interesting insights: The regularity of the world we observe around us is extremely improbable. The process of reducing a set of observations to a general theory explaining these observations can be described as a process of data-compression. A universal methodology of science would have the following form:

- Represent your data set D in binary format.
- Select a hypothesis h in binary format such that $|S(h) + S(D|h)|$ is minimal.

This program fails because of the uncomputability of S but it can serve as a regulative ideal for the study of methodology of science. In certain cases the theoretical results allow us to solve real life problems and to develop more efficient algorithms (Li and Vitányi [1997]). Note that we have characterized learnable data sets as non- and sub-extensive, they contain a mix of random and deterministic elements. MDL aims at finding a compression for such a set that exactly separates the random (extensive) elements ($S(D|h)$) from the non-random (intensive) ones ($S(h)$). For intensive data sets the two part code will simply consist of a description of the program generating the data set ($S(h)$) and the length of the data set ($S(D|h)$).

Another way to look at this is from the perspective of the so-called *randomness deficiency*(Vereshchagin and Vitányi [2004]):

$$\delta(D|M,d) = \log \binom{m}{d} - K(D|M,d), \tag{1}$$

Here M is a model of size m and $D \subseteq M$ is a data set of size d. The expression $\log \binom{m}{d}$ is the measure of the maximum entropy of a subset of M of size d. The expression $K(D|M,d)$ is the actual entropy of the data set D in the model, i.e. conditional Kolmogorov complexity of D given M and d. If the actual entropy is much smaller than the maximal entropy of an average set of size d in M then D still contains a lot of regularity that is not explained by M. In other words M is not an optimal model. A model would be optimal if the randomness deficiency is minimal. In such a case D would be a typical element (extensive) of M and M would explain all that is worth knowing about D, i.e. its intension. The principle of minimal randomness deficiency is very close to Jaynes'maximal entropy principle: in order to explain a set D try to find the set M for which the entropy is maximal under a set of constraints observed in D. [21]

5 The Cooperative Computational Universe

From this discussion it is clear that the philosophy of learning touches on a number of philosophical issues: To name a few: entropy, information, computation, objective and subjective probability. In order to study these issues let's define a thought experiment. For the sake of argument we will restrict ourselves to the case in which we observe a string of bits from an unknown source. Even in this simple setting there are some fundamental philosophical issues to be dealt with.

Suppose that we reserve a room at the university of Amsterdam for the purpose of this experiment. The room has no windows and the door is closed. In the room there is a black box. The black box produces a bit every minute. If the bit is '1' the light is switched on, if it is '0' the light is switched off. This bit is published on a web site. Of course nobody knows the contents of the black box, but for the sake of arguments we choose three possible configurations. The box could contain:

[21] See the paper of Bais and Farmer in this book.

1. A *random process* that generates bits (e.g. a person flipping a coin, or some other ergodic process.).

2. A *deterministic computer program* generating bits.

3. An *infinite database* with a list of bits.

These three definitions represent radically different views on the phenomenon of a source of information. The first is an objective random process associated with an objective form of probability. It generates an extensive data set. All the information that is contained in the sequence can be measured in terms of its fundamental statistical characteristics: mean, variance, autocorrelation function etc. The second is a deterministic process with a definition of finite length. The maximal amount of information in a string produced by the program is limited to the length of the definition of the program. It is an intensive data set. It could lead to a sequence of bits with a certain statistical bias (e.g. repeating patterns), but this is not necessary. Some transcendental numbers have short definitions (e.g. e and π) but lead after a bit of twisting to bit patterns that cannot be recognized as non-random. The third is a deterministic process with a definition of infinite length. The generating data set itself is could be in- or extensive. It potentially contains an infinite amount of information that can never be learned in a finite amount of time.

Theorem 5.1 *The three sources of information, (a random process, a deterministic computer program and an infinite database) cannot be distinguished from each other by a receiver of the information.*

Proof: Each of the three sources can produce a sequence of bits that cannot be distinguished from a random sequence. 1) The case of the random process is trivial 2) A deterministic program can generate strings that cannot be recognized as non-random. The non-computability of Kolmogorov complexity tells us that there will always be compressible strings for which no compression can be computed. 3) An infinite database can continue a random set of bits or a set of non-random bits that cannot be recognized as such.

The philosophical importance of this result is obvious. We cannot make a distinction between a source of information that is random and a source of information that has high complexity. This makes the traditional controversy between determinism and indeterminism from the point of view of informatics senseless. It reveals the famous dictum by Einstein "God does not play dice" as a real metaphysical position. It is not a question that can be settled by any argument. It also shows that it is impossible to assign any form of objective probability to a source of information. In this context one might ask to which extent randomness is in any sense a scientific concept. We can define randomness of strings in terms of incompressibility, but we do not need the concept of randomness to study incompressibility. The notion of flipping a coin or throwing a dice are real scientific paradigms in the original Kuhnian sense, but au fond they are deterministic processes that in most cases are simply too complex to

predict and therefore can act as place holders for supposedly real random processes. They serve as anecdotic topoi in the scientific discourse, nothing more. The notions of extensiveness and incompressibility still have an exact meaning in a deterministic Laplacian universe, so they seem to be more fundamental than the concept of randomness. Macroscopic measurements of microscopic deterministic processes might subjectively be interpreted as random. Even in a Laplacian universe there are data sets that are both strictly deterministic and extensive (e.g. the Halting set).

In such a world however there is a form of subjective probability that is relevant. Suppose that we want to form a hypothesis about the internal structure of the black box and the black box produces a string that shows some regularity. In that case it is extremely unlikely that the source of bits is random. Suppose that our black box produces a string of n ones $1_1 1_2 \ldots 1_n$. The probability of creating this string with n flips of a perfect coin is 2^{-n}. So, intuitively, with each one that is produced by our black box the hypothesis that it contains a random process becomes more unlikely in favor of the hypothesis that the bits are produced by some deterministic process. Yet this argument is flawed because *any* bit string of length n produced by flipping a perfect coin has probability 2^{-n} and therefore is extremely unlikely. We have no clear ground to favor any regular string over a random one as a ground for selecting between hypotheses about the content of the black box. As we have seen, the theory of Kolmogorov complexity allows us to define the concept of *randomness deficiency* of a string. The idea is the following. A string like, say, 11100101000100 is *typical* for a random source. Such a string is produced by a source that is perfectly compatible with the hypothesis that the source is random. A string like 11111111111111 is *atypical* for a random source. When produced by a source it makes the hypothesis that the source is random unlikely. A high randomness deficiency corroborates the theory that the process in the black box is non-random.

This analysis suggests that the best thing we can do in science is: observe a set of phenomena, estimate the randomness deficiency and formulate a theory. Unfortunately in the case of the Amsterdam room the situation is more complicated. This becomes clear if we analyze the following claims.

Claim 5.2 *We get exactly one bit of objective information each minute.*

It is clear that each bit that is published on the web by the black box contains real information about the actual binary situation in the room: the light is on or off.

Claim 5.3 *The meaning of the message contained in the bit and the knowledge generated as a consequence of receiving the message is not dependent on the content of the black box.*

Yet there is a subtle interplay between the growth of our subjective information and our theories about the nature of the black box.

Claim 5.4 *The objective amount of information we get is dependent upon our interpretation of the nature of the source of information.*

The three possible interpretations of the content of the box could be seen as three different types of senders of messages. I will define three possible receivers along the same line:

1. A forgetful receiver that determines the statistical characteristics of the sequence: mean, variance, autocorrelation function etc. Here our subjective information grows incrementally at a very slow rate with each objective bit that is received. This observer corresponds with an interpretation of the source as a system in equilibrium. The statistical (macroscopic) qualities of the system are all that we can know about the system.

2. A machine learning program with bounded computing time and memory, that tries to reconstruct the finite structure of the black box. Here our subjective information grows in an irregular but monotone way with each bit of objective information that is received. This observer corresponds to an interpretation of the data set as intensive. After some finite point in time our information will only grow with the factor $logx$ where x is the number of bits we have seen so far.

3. An infinite database with a list of bits recording every bit that is received. Here our subjective information grows with exactly 1 bit per bit that is received, if the data set itself is considered to be extensive.

This example shows that we can not restrict ourselves to a purely subjective interpretation of information when we analyze a source of messages. We need to make an a priori decision about the nature of our source.

Our analysis shows that nature and science play an asymmetrical game. Non-random strings are very rare. To make this more specific: in the limit the density of compressible strings x in the set $\{0,1\}^{\leq k}$ for which we have $K(x) < |x|$ is zero. Data sets that appear to be random may be actually compressible, but the occurrence of such objects in nature is extremely unlikely. If a data set looks random, we may with high probability assume that it is random. On the other hand if a data set from the point of view of an intelligent agent appears to be regular then it is with extremely high probability not random and can be learned because of the shallowness claim 4.12. Therefore a learning system that simply scans the environment for areas of low entropy and tries to compress the data sets it finds there will be successful with high probability, if the complexity of data sets is of the same order of magnitude as the agent. Local low entropy data sets correspond with energy consuming non-equilibrium systems that with high probability can be described in terms of computational models. Learning is not as hopeless as our formal models seem to imply. We are computational processes of limited complexity analyzing computational processes of limited complexity in a universe that generates computational processes of limited complexity. In this sense we live in a cooperative computational universe. This is as close as we can get to the solution of certain philosophical problems in terms of information and computer science.

So why is this the case? Why do we live in a world that is intelligible at all? This question pervades philosophy from its early conception on (Herakleitos vs

Parmenides). In form of a sweeping statement: prima facie, the god of Leibniz might very well have created a universe in which the Minimum Description Length principle would not hold. There seems to be no theoretical necessity to favor simplicity. The extreme regularity of the universe could be a 'local' condition accidentally observed by us. In terms of modern information theory: every infinite random string has an infinite number of regions of extreme regularity. If we transpose this idea to the analysis of our world we might just accidentally live in such a regular region in a purely random universe. Li and Vitányi [1992] A rather horrifying thought.

On the other hand imagine the following thought experiment: an infinite set of universal Turing machines working in parallel with input tapes that are created by means of some random process (e.g. flipping a coin). The set of input tapes is infinite so every finite prefix free program will occur an infinite number of times. Yet the density of 'shorter' programs will be exponentially higher than that of 'longer' ones. Some programs will run for ever, others will stop in finite time. After n time steps a number of 'simple' programs will have stopped and produced a fixed output. This means that the set of outputs we observe in this thought experiment will have a strong bias for simplicity. In other words even a universe that consists of purely random computational processes has a strong bias for simplicity. The distribution of phenomena it produces is cooperative in the sense that we get examples of the simple structures first. This is the hypothesis of the cooperative universe in another guise: nature produces the information that we need to interpret her in such a way that hypotheses we form are right with high probability. In such a universe MDL therefore will be a viable methodological principle. It coincides with another well known dictum of Einstein: Subtle is the Lord, but malicious He is not. The exact relation between various computational models of the universe, cooperative distributions, the universal distribution $\mathbf{m}$ and the problem of induction is, in my view, one of the most important open problems in the philosophy of information.

These issues (subjective versus objective probability, regularity versus randomness, information versus meaning) are far from resolved and should be at the center of a philosophical research program of a philosophy of information.

6 Conclusion

The research on learning and induction that has emerged because of the growing interest in artificial intelligence is still developing. The results do not only lead to useful industrial applications, but also influence the way we think about fundamental philosophical questions about the origin of human knowledge, the structure of our brain and methodology of science. A formal analysis of the mathematics of learning helps us to understand the efficiency of human learning. Human beings can only learn complex structure like language and the laws of nature if the underlying probabilities are 'benign'. The hypothesis of the cooperative universe is an attempt to explain why we live in a world that can be learned efficiently.

Finally a tongue in cheek observation. Our human brain can contain about 10^{14} bits of information. The total storage capacity of the known universe is estimated to be about 10^{92} bits (Lloyd and Ng [2004]). The old philosophical ambition of understanding the universe as a whole amounts to the wish to find a compression of the universe of the following nature: a structural description of less than 10^{14} bits (the laws of nature) and an ad hoc description of more than 10^{78} bits (the actual structure given the laws of nature) . There is only one conclusion possible. The universe can only be understood by human beings if it is extremely compressible -in other words- if almost nothing of any significance happens.

References

Adriaans, P. (2001). Learning shallow context-free languages under simple distributions. In Copestake, A. and (eds.), K. V., editors, *Algebras, Diagrams and Decisions in Language, Logic and Computation*. CSLI/CUP.

Adriaans, P. and Vitányi, P. (2005). The power and perils of MDL. Technical report, Human Computer Studies Lab, Universiteit van Amsterdam.

Adriaans, P. and Zantinge, D. (1997). *Data mining*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Adriaans, P. W. and van Zaanen, M. M. (2004). Computational grammar induction for linguists. *Grammars*, 7:57–68.

Angluin, D. (1988). Queries and concept learning. *Machine Learning*.

Black, M. (1967). Probability. *The Encyclopedia of Philosophy, Paul Edwards (ed.)*, 6:464–479.

Capurro, R. (1978). *Information. Ein Beitrag zur etymologischen und ideengeschichtlichen Begründung des Informationsbegriffs*. München, New York, London, Paris: Saur.

Capurro, R. and Hjørland, B. (2003). The concept of information. *Annual Review of Information Science and Technology*, 37(8):343–411.

Carnap, R. (1950). *Logical foundations of probability*. The University of Chicago Press.

Cilibrasi, R. and Vitanyi, P. (2005). Clustering by compression. *IEEE Transactions on Information Theory*.

Cornuéjols, A. and Miclet, L. (2003). *Apprentissage artificial, concepts et algorithmes*. Eyrolles.

Domingos, P. (1998). Occam's two razors: The sharp and the blunt. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*. AAAI Press.

Domingos, P. (1999). The role of occam's razor in knowledge discovery. *Data Mining and Knowledge Discovery*, 3(4):409–425.

Dretske, F. (1981). *Knowledge and the Flow of Information*. Cambridge MA: MIT Press.

Edwards, P. (1967). *The Encyclopedia of Philosophy*. Macmillan Publishing Company.

Floridi, L. (2004). Open problems in the philosophy of information. *Metaphilosophy*.

Hájek, A. (2002). Interpretations of probability: http://plato.stanford.edu/entries/probability-interpret/. Stanford Encyclopedia of Philosophy, ed. E. Zalta.

Hawking, S. (1988). *A brief history of time: from the big bang to black holes*. Toronto; New York: Bantam Books.

Hume, D. (1909, 1914). *An Enquiry Concerning Human Understanding*, volume Vol. XXXVII, Part 3 of *The Harvard Classics*. P.F. Collier & Son.

Kearns, M. and Vazirani, U. (1994). *An introduction to computational learning theory*.

Kneale, W. and Kneale, M. (1988). *The Development of Logic*. Oxford, Clarendon Press.

Li, M. and Vitányi, P. (1997). *An introduction to Kolmogorov complexity and its applications*. Springer-Verlag, 2 edition.

Li, M. and Vitányi, P. M. B. (1992). Philosophical issues in Kolmogorov complexity. *Automata, Languages and Programming: Proc. of the 19th International Colloquium*, pages 1–15.

Lloyd, S. and Ng, Y. (2004). Black hole computers. *Scientific American*.

Locke, J. (1961). *An Essay Concerning Human Understanding*. London : Dent ; New York : Dutton.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill, New York.

Popper, K. (1952). *The Logic of Scientific Discovery*. London: Hutchinson & Co. Postman, L., & Brown, D.R.

Rissanen, J. (1999). Hypothesis selection and testing by the MDL principle. *The Computer Journal*, 42:60–269.

Solomonoff, R. J. (1997). The discovery of algorithmic probability. *Journal of Computer and System Sciences*, 55(1):73–88.

Solomonoff, R. J. (2003). The Kolmogorov lecture. The universal distribution and machine learning. *Computer Journal*, 46(6):598–601.

van Aartrijk, M. L., Tagliola, C. P., and Adriaans, P. W. (2002). AI on the ocean: the robosail project. In van Harmelen, F., editor, *ECAI*, pages 653–657. IOS Press.

Vereshchagin, N. and Vitányi, P. (2004). Kolmogorov's structure functions and model selection. *IEEE Trans. Information Theory*, 50(12):3265–3290.

Weaver, W. and Shannon, C. (1949). *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, Illinois. republished in paperback 1963.

Windelband, W. (1921). *Lehrbuch de Geschichte de Philosophie*. Tübingen, Verlag von J.C.B. Mohr (Paul Siebeck), 9,10 edition.

Wolfram, S. (2001). *A new Kind of Science*. Wolfram Media Inc.