Philosophy of Information, Concepts and History
Preliminary Version

version 1.0 References need to be added and checked

Pieter Adriaans (HCS, Amsterdam).

Abstract

In this paper I present an outline of a philosophy of information from a systematic and a historical point of view. In the first part I discuss the concept of a Turing machine and various concepts of information, mainly focusing on Shannon information and Kolmogorov complexity. I show that these concepts can be interpreted as mathematical guises of a common notion of information that is associated with the entropy of data sets. Issues concerning the methodology of science, optimal coding, data compression and induction are discussed in this context. In the second part of the paper I show that these notions are rooted in the history of philosophy and that a philosophy of information can be seen as the resolution of an age old philosophical ambition to create a universal language of science.

1 Introduction

According to the legend Theseus forgot to change the sail of his ship from black to white on his return from Crete. He had promised his father to do this as a sign of the fact that he had successfully defeated the Minotaur. The old Aegeus, standing on the lookout, thought his son was dead and threw himself from the cliffs. Any computer scientist could have pointed out that it is dangerous to use a non-redundant one bit coding scheme for such an important message. This example shows that our thinking about information encoded in bits has evolved far beyond the traditional application in computer programs. Information seems to be everywhere and almost everything seems to be associated with information. The notion of information is a key category of modern science. In this paper I will examine some of the roots of the concept of information and its relevance for philosophy.

The central motivation for the philosophical study of information, in my view, is the fact that the concept of information with its embedding in a fundamental mathematical framework is the closest we can get at this moment in history to the realization of two very old philosophical ambitions:

- A unified mathematical description of reality and
- A unified scientific language.

I do not claim that these two ambitions are completely resolved, nor that philosophical reflection on information is in any way finished, but I do claim that recent insights in this domain shed new light on older philosophical research programs and in some cases allow us to revitalize them. It is the aim of this paper to sketch the contours of such a philosophy of information that can be defined as:

- The philosophical analysis of the concepts of 'information' and 'computation'.
- The philosophical analysis of the role of information in nature, science and culture.
- The analysis of the notion information in the context of traditional philosophical disciplines like metaphysics, methodology of science, epistemology, ontology, ethics and aesthetics.

Philosophy of information is a young discipline with unclear boundaries. A valuable attempt to define the subject is the paper of Floridi in this book, who lists 18 open problems (See also: FLoridi [2004]). If anything, this list shows that philosophy of information is far from mature. The length of the list is rather depressing. The internal relation between the various problems is unclear and the importance of the problems varies. The ambition of this paper is to present a coherent view on the subject. I am well aware of the fact that this view is debatable and that other approaches to the philosophy of information are possible, but I think that the value my contribution lies in the fact that it can be a starting point for further study and discussion.

There is no ambition to be comprehensive. One of the major problems in writing this paper was selection of the material, especially for the historical paragraphs. There are a lot of relevant sources that I do not mention because of lack of space, personal preference, or possibly because I was not aware of their existence. What is presented could be compared a bus tour through the domain, pointing out interesting locations for further study. Given the present state of the subject I think this ad hoc approach is justified. The paper is intended as a sketch of a research program that I hope will take shape in the coming decades.

This paper is organized in three sections. I start with a description of the current views on information. Then I continue with a section on history where I show how the concept of information in the context of notions of a mathesis universalis and a unified language for science is deeply rooted in the history of philosophy. In the last section I analyze some of the central problems in the current philosophical reflection on the notion of information.

2 Concepts

'Information' is a term that is much like 'energy', 'entropy' or 'force'. It has half a dozen or so precise mathematical definitions and an almost unlimited number of colloquial meanings. The starting point of our investigation will be the mathematical concepts of information as they are treated in basic introductory texts in computer science (e.g. Hopcroft-Ullman, Cover, Li -Vitányi). Some competing notions of information are:

- Information as a common sense concept.
- Shannon information: bits: $I(x) = -log_2 P(x)$.
- Algorithmic information: $K_U(x|y) = min\{|p| : p \in \{0,1\}^*, U(p,y) = x\}$.
- Fisher information: $I(p) = \int \frac{[p'(x)]^2}{p(x)} dx$.
- Quantum information stored in qubits

All of these concepts, except Fisher information, are treated extensively in this volume. Later we will see that the historical background and the cultural impact of these mathematical definitions for science and for every day life are vast and highly non-trivial, but for the moment we will be naive and stick to standard formal definitions. These notions are associated with a number of very powerful ideas that are crucial for information theory:

- The mathematical definition of the concept information in terms of the probability of a message.
- The definition of the bit as fundamental unit of information. The bit is defined as the maximal amount of information that one can obtain from a yes/no question (See Harremoës and Topsøe in this book).
- The association of mathematical proof with computation as a sequence of well-defined events in the physical world.
- The definition of the a priori probability of a binary object in terms of its computational complexity.

The concept of a message sent from a sender to a receiver can be seen as a true paradigm of modern science in the Kuhnian sense, just as the notion of a perfect collision between bodies is a fundamental paradigm of Newtonian physics and the notion of the primal scream (Uhrschrei) is a founding concept of romantic thought. If one tries to deconstruct the modern notion of information the following elements seem to come in to play:

- An underlying *transaction*. The notion that information *flows* between the sender and the receiver and the notion that the information of the receiver *grows* as a result of reception of the message.
- A *code system*. The notion that messages can be coded in terms of *systems of arbitrary signs*.
- A *mathematical measure of information content* of the message.

One of the philosophical conceptions underlying this *information paradigm* is a radical re-interpretation of the importance of language that took shape in the twentieth century. Anything that science can say about the world has to be expressed in language and therefore the starting point of any philosophical reflection should be an analysis of language. The history of this development is well-known and can be followed in any adequate textbook on philosophy. Traces of these views can be found in the work of philosophers as diverse as Boole, Frege, Husserl, Russell, Wittgenstein, Carnap, Heidegger, Feyerabend, Popper, Lakatos, Searle, Austin and Derrida. An emerging philosophy of information builds on these developments. It would, in my view, be wrong to interpret the task of philosophy of information as a mere continuation of these ideas. Its claim to fame is that it brings a form of mathematical rigor to the discussion that carries the promise of philosophy as a real foundational discipline. It sheds new light on a number of central philosophical problems, not only in the domains of philosophy of knowledge and methodology of science but possibly even in ethics, esthetics.

The universal Turing machine and a universal language of science

There are various possible notions of the concept of computation. To name a few:

- Pythagoras' model: addition, substraction, multiplication, division.
- The Gödel model: recursive functions.
- The Turing model: Turing machines.
- The Church model: the lambda calculus.
- The Wolfram model: cellular automata.
- The Quantum model: quantum computing.

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 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 2.1 *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 2.2 *Any finite discrete system or process can be described in terms of operations on natural numbers.*

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 [1948]. Here the information content of a message is defined in terms of its probability:

Definition 2.3 *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[1]: 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 fill this gap is made Dretske (1981, also this book !!!). 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.

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 this 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:

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

[1] *log* is used for *log*$_2$

. 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. [2] 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 (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 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 are 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 the 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

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

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

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 $K_U(x|y) \leq K_T(x|y) + c_T$ [3]. 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 2.4 *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.*

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

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 [4]. 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 exist 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 0101010101010101010101010101. We can *recode* this string in Shannon's sense as '01'=1;11111111111111, or 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 2.5 *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 2.6 *The concepts of Kolmogorov complexity and Shannon information are equivalent in the case of data sets with maximal entropy.*

[4] For a more extensive discussion of these issues, see the chapter by Grüwald and Vitányi in this book.

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.

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. Using results of Solomonoff Solomonoff [1997, 2003] and Levin we can define an a priori probability of a finite binary string.

Definition 2.7 (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 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 2.8 (Levin)

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

The philosophical importance of these concepts can not be overstated. They offer new general solutions for age old problems. The universal distribution has quite wonderful qualities and its philosophical relevance has hardly been explored up till now.

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 $S : \{0,1\}^* \rightarrow \{0,1\}^*$ as:

Definition 2.9 $S(x) = \min_{i \in N}\{p : b(i,p), U(p,\epsilon) = x\}$

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 2.10 *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 'nearly' 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.

Let us have a closer look 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 2.11 *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. 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 2.12 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 (!!! ref Domingos). 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 that h_i is better than h_q. This can only be guaranteed if $S = S_i = S_p$, i.e. when the code is optimal (ref !!! Adriaans and Vitányi, forthcoming).

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 as a regulative ideal for the study of methodology of science. In certain cases the theoretical results allow us to solve real life problems (ref !!! incompressibility method) and to develop more efficient algorithms.

3 Historical roots of the concept of information

One can safely say that the explicit abstract notion of information as it was introduced in twentieth century computer science was absent in antiquity. When we read the pre-Socratic philosophers like Zeno, Parmenides or Plato with our modern mind, we feel uneasy about the undifferentiated mix of formal, epistemological, ontological, ethical and esthetical questions. It is all there, but without the distinctions. The same holds for the notion of information. With hindsight one could say it has played a role in philosophy from the beginning, without being recognized as such. The history of the concept of information is related to, but should not be identified with: the history of the term 'information', the history of the computer, the history of logic or the history of epistemology. The intention of the following paragraphs is not to give a full fledged history of information, but more to point at interesting ancestors of the modern approach.

A note on the history of the term 'information'

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 (!!! ref) and Augustine seems to have played no role in the emergence of the modern concept of information. In the 15th century the French term 'information' finds its way into the colloquial vocabulary of European languages with various subtle difference 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. Even Edward's Encyclopedia of Philosophy from 1967 does not have a separate lemma on information. [ed.] The same holds for the well-known History of Logic written by Kneale and Kneale that first appeared in 1962. In short the term 'information' seems to have been absent from the philosophical dialogue for a couple of hundred years.

The history of the emergence of the technical term 'information' in the 19th and 20th century has yet to be written. It appears however that it is closely connected with the rise of modern intelligence services and the development of new means of communication like the telegraph. At the end of the 19th century several countries created departments with the responsibility to collect military information. In 1866 the German government started a Foreign Office Political Field Police with the mission to procure intelligence about the Austrian enemy army. After the Franco-German war in 1870 the French created a *Statistical* and Military Reconnaissance Section (italics mine) with the specific task to collect information on German military operations. In 1873 the British War Office established an Intelligence Branch, staffed by twenty-seven military and civilian personnel. In the USA a Office of Naval Intelligence was established in 1882 followed by a Military *Information* Division (MID, italics mine) -with one clerk and one officer- in 1885. Its task was to collect "military data on our own and foreign services which would be available for the use of the War Department and the Army at large." (see: !!! A Century of Spies by Jeffrey T. Richelson). The enormous influence of intelligence services in the 20th century is well-known. An example of modern general use of the term information in this context can be found in the World Fact Book, an annual publication of the CIA: *Information is raw data from any source, data that may be fragmentary, contradictory, unreliable, ambiguous, deceptive, or wrong. Intelligence is information that has been collected, integrated, evaluated, analyzed, and interpreted.* The shift in meaning of the term 'information' from 'information as the act of informing' to the 'information as the result of the act of informing' to 'information as something that is contained in the message that is used to inform' is striking and relevant.

Code systems

Coding systems are as old as language itself. The ambition to hide information in messages and to decode these messages without knowing the key has always existed. (Ref Kahn !!!). Ceasar already used code systems to communicate with his generals. Technically one has to distinguish various techniques: *Ciphers* use letter transposition and substitution systems (Ceasar alphabet, cipher disks, Vigenère tablaux, Enigma). *Codes* consist of lists of codewords and code numbers. The idea that it is efficient to assign the shortest codes to the most frequent signals was known long before Shannon defined its mathematical basis in 1948. With the invention of bookprinting in the 15th century typefounders directly made the empirical discovery that they needed more e's than z's in a font. The fact that the frequency of specific letters in a text is typical for a certain language was well-known and was used to decode simple letter replacement ciphers. The 18th century saw the emergence of so-called blackrooms in Europe with the sole task to encode and decode messages for political purposes. Around this time the first elementary statistical techniques to decipher text on the basis of frequency analysis were developed. With the development of the first electronic communication media the question of efficient coding systems became urgent. In 1838 Samuel Morse designed his famous code on the basis of a statistical analysis of the number of letters in the typecase of a Philadelphia newspaper. The most frequent letter 'e' got the shortest symbol, a dot. The next frequent letter 't' was encoded as a dash. The publication of Shannon's paper also marked a breakthrough in our understanding of code systems. Almost ciphers and codes can be cracked with statistical techniques, given enough data. An exception form the so-called one-time-pad ciphers (Shannon [1948]).

Entropy

From thermodynamics to quantum mechanics the study of physics is covered with deep questions concerning the nature of information and our capacity to know the world around us. The confrontation between physics and information theory often lead to important breakthroughs in both disciplines. An extensive discussion of these issues is found in the chapter by Bais and Farmer in this book.

The idea that matter consists of elementary particles dates back to Democritus. It was restated in 1802 by the English school teacher John Dalton. The notion of atoms immediately leads to interesting epistemological problems. Suppose that we want to analyze the behavior of a certain amount of gas in a closed container. The gas has a certain temperature and a pressure. If we reduce the volume of the container the statistical probability of collision between a particle and the wall will increase. On a macroscopic level this will be observed as an increase of pressure. James Clerk Maxwell was the first to introduce probabilistic methods into physics in this context [!!! "Illustrations of the Dynamical Theory of Gases," 1860].

The explanation of macroscopic continuous events in terms of a statistical analysis of large quantities of discontinuous microscopic events leads to epistemological questions. In a container with only a handful of particles the concept of pressure has no meaning. It only exists in the presence of extremely large quantities of elementary particles. Al things being equal we observe the pressure of the gas to be constant at the macroscopic level. At the same time at the microscopic level there is a constant change of configurations. A macroscopic phenomenon like pressure is called emergent. It does not exist at the microscopic level. We can reason about the macroscopic events without any information about the specific microscopic configurations. This is striking. Configurations in which all of the particles are in one half of the container are highly unlikely, but not logically impossible. In such a case the pressure in one part of the container would be zero and twice as big as the original pressure in the other half. If the container would be a cylinder and we place a piston in it then the gas would push the piston to one side of the cylinder. If the pressure on both sides of the cylinder is equal no movement will be observed. Apparently some configurations can be converted into kinetic energy. Others lack this potential.

In order to explain these phenomena Clausius introduced the concept of entropy. Clausius [1850] Entropy is a measure of the total number of different microscopic states the macroscopic system can exist in. The entropy in the container is higher if the particles are evenly distributed over the space in the container. With the concept of entropy Clausius could formulate what later has become known as the second law of thermodynamics: a closed system will remain the same or become more disordered over time, i.e. its entropy will always increase. The philosopher Henri Bergson called this insight "the most metaphysical law of nature" (Henri Bergson, Creative Evolution !!!). Clausius ended his paper with a rather disturbing observation: "The energy of the universe is constant-the entropy of the universe tends toward a maximum."

The importance of thermodynamics for information theory can hardly be overstated. There is an striking equivalence between the mathematical formalization of the concepts of entropy and Shannon information. The total entropy of two independent systems is the sum of the individual entropies while the total probability is the product of the individual probabilities. Boltzmann therefore proposed to that the entropy of a system would be proportional to the logarithm of the number of microstates a system could be in. Likewise Shannon proposed to express the amount of information in a message in terms of the base two logarithm of its probability. The total information in two independent messages is the sum of the individual information per message. The probability of two messages occurring together is the product of their individual probabilities. The use of the base two logarithm ensures that the information in a message can be expressed in bits. Shannon also introduced the concept of entropy of a set of messages. The entropy is maximal if all the messages in the set have equal probability. A related notion of entropy can be defined in relation to Kolmogorov complexity. Strings with low Kolmogorov complexity have low entropy. Random strings have high entropy. The mathematical kinship between thermodynamics and theory of information ensures an almost seamless translation of concepts of one theory to the other. In sweeping statement one might say that information theory is the thermodynamics of binary strings while thermodynamics is the information theory of particles in space.

Recently physicists have taken this analogy to the extreme by analyzing black holes and even the whole universe as a computational system. Lloyd and Ng [2004] The aptness of this analogy is doubtful. If one defines computing as 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, then it is not directly clear how a kilo of pure plasma or a black hole could be interpreted as a computer. These conceptions rather seem to be associated with localized parallel random processes that are only in a very specific sense equivalent to computing. There is certainly no stable data storage in a kilo of pure plasma although theoretically it can store 10^{30} bits of information.

The minimum description length principle (MDL) and induction

MDL is often related 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.

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 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 two 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] He showed that the algorithmic or Kolmogorov complexity of an object is associated with an a priori probability of this object. The impact of this insight can hardly be overstated. It allows us in theory to assign an 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 unified description of Nature

I have interpreted the philosophy of information as the resolution of two old philosophical ambitions:

- A unified mathematical description of reality and
- A unified scientific language.

As is to be expected the conceptualization of these notions in philosophy is a complex process that could easily fill a monograph. It was a cumbersome development that left traces throughout the history of philosophy. But the roots of the concept of information are certainly not predominantly philosophical. The application of mathematics to different regions of reality has traditionally been the domain of the engineer. In this sense there is an fascinating dialectic between the bold statements of early philosophers and the meticulous process of the mathematical conceptualization of reality in the history of science. A birds eye view of this development would show the following rough phases: ca. 600 BC: space, music, ca. 1500 AD: seeing, light, ca. 1600: temperature, movement, ca. 1700: force, probability, ca. 1850: thermodynamics, ca. 1900: language: ca. 1940: social/information, ca. 1970: chaos/complexity.

The earliest traces of the thought that the essence of the world is mathematical are found in the writings of the Pythagoreans. The later Plato seems to have cherished similar thoughts. In the Politicus he refers to them and states that everything in the world and in human life is object of geometry (Polit. 285a). In his Metaphysics Aristotle repeatedly mentions Plato's view that the ideas are essentially numbers. What Plato exactly had in mind is unclear and still a matter of debate. He seems to have limited to the amount of 'idea-numbers' to ten and his thoughts were probably more mystical than related to any modern mathematical views, but the idea was powerful and had a decisive influence in history.

It has also bearing on the notion of a universal language of nature. The ambition formulate ideas clearly has been to prime goal for philosophers from the early start in Greek thought. With the rise of modern science the need for unequivocal terminology and efficient language became apparent. Plato was well aware of the problem of the conventional nature of natural language. It is discussed extensively in the Cratylus. If language is purely conventional then there is no guarantee that our linguistic concepts are scientifically adequate. The challenge for philosophy then would be to construct a pure natural language that by nature would be adequate to express scientific concepts. Also in Aristotle's writings (Specifically De Interpretatione and the Categories) this ambition is present. They would remain a predominant influence throughout history although we have to wait till the end of the 19th century before the notion the mathematical study of language as a purely arbitrary system of signs emerged. During the middle ages the study of language and logic as a tool for philosophy reached unprecedented heights culminating, via influences of Porphyri, Boethius, in the Dialectica of Abélard (1079-1142) and Ockham's theory of suppositions. Medieval mathematical thought is underdeveloped and the philosophical investigations mainly focus on the relation between language and logic. But also in this context we find powerful ideas about the intricate relation between the structure of language and the world. Early Christian philosophers proposed the doctrine that god is the author of two texts: the Bible and the book of nature. Understanding nature is equivalent to deciphering the book of nature.

"*For this whole visible world is a book written by the finger of God, that is, created by divine power [...] But just as some illiterate man who sees an open book looks at the figures but does not recognize the letters: just so the foolish natural man who does not perceive the things of God outwardly in these visible creatures the appearances but does not inwardly understand the reason.*" (12th century Hugh of St. Victor qtd. in Josipovici 29)

The thought that nature is a text that needs to be deciphered plays an important role in modern philosophy (Foucault, Derrida) as well as in every day life. It gives the concept of the computer as a universal information processing machine a natural embedding much older and general philosophical reflection.

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 (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 the 17th century is a central period in history of the conceptualization of information. Descartes 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 Descartes man 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 is problematic. With hindsight one could say that in the work Descartes the need for *an abstract concept of mediation between knower and the known*, i.e. a concept of information, became problematic for the first time. Because of this lack, he was incapable of developing an adequate philosophical theory of language.

The first 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). The carthesion cogito becomes a epistemological subject that starts as a tabula rasa and is gradually filled up with 'ideas' that find their origin in experience. Lock 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. Next to Pysika and Praktika Locke considered the study of Semiotika to be on of the three cornerstones of science: "*the business whereof is to consider the nature of signs the mind makes use of for the understanding of things*". (Essay, IV, xxi,4). In part III of the essay Locke develops a theory of language. The philosophical value of this part of his work is limited but the historical influence on the 18th century philosophical conception of language has been enormous.

There have been numerous proposals for artificial languages that would somehow make scientific communication more efficient and or reliable (Bacon, Wilkins, Leibniz, to name a few). Around the same time that bishop Wilkins was conceiving his artificial language in England the grammarians of Port Royal in France developed their universal grammar. The first author to propose a universal character was Francis Bacon but the basic strategy of every proposal is the same: step 1) develop an ontology, step 2) Assign a sound or a visual symbol to each basic concept in the ontology, step 3) design a syntax to combine the symbols. There are two distinctive features that separate these proposals from the modern conception of a universal language. The first is, what one could call, the ontological assumption (cf. step 1). None of the authors doubted that an ideal language for science with an ideal set of concepts was in fact possible. The second difference between early artificial languages and the modern conception of a universal language is the concept of a purely formal language that is completely defined in terms of its syntactic operations. It was developed in the second half of the 19th century. (George Boole: An Investigation of the Laws of Thought, London 1854, Gottlob Frege: Begriffschrift, 1879, Russell-Whitehead;

Principia Mathematica, 1910).

Apart from coining neologisms scientists followed two strategies to identify the ideal language for science: they declared an existing language (Latin or their own) to be superior to any other for this purpose or they designed an artificial language. Simon Stevin for example believed the Dutch language to be ideal for science because of its tendency to denote 'single things with single sounds' (ynckel saken met ynckel; gheluyden te beteeckenen. Uytspraeck vande Weerdicheyt der Duytsche Tael 1586). It is instructive to note that Stevin tried to prove his point by making list of single syllable words for various languages. In fact he argues that the Dutch language is a near optimal coding system for scientific concepts. Throughout the 18th century we seen an explosion of essays touching on the structure and origin of language. In England Warburton and Monboddo, in Italy Giambattista Vico, in France Maupertius, Condillac, Rousseau, Diderot, l'Epée and Condorcet, in Germany Leibniz, Wolff, Süssmilch. The thought that language is more than a neutral tool and can be a barrier for communication is of later date (the Romantic period: Hamann, Herder, Von Huboldt, Sapir-Whorff). It is foreshadowed in Rousseau's Essay sur l'origine des langues (Ref!!! See: La tranparance et l'Obstacle by Starobinski). The stream of publications on language and philosophy swells in to a river in the 19th century. The notion that understanding the universe involves calculation in a universal language is a central idea in Leibniz' Characteristica Universalis. The first person to realize that a mechanical calculator could be used as a universal information processing machine seems to have been lady Ada Lovelace working with Charles Babbage on the so-called Analytical Machine in 1842. The notion culminates in Turing's definition of a universal machine.

The true precursor of a modern philosophy of information is the research in to the foundations of mathematics starting at the end of the 19th century. Hilbert had formulated his formalist program for mathematics, which is the culmination of the ambition implied by Locke's Semiotika: the study of mathematics as pure formal manipulation of signs. Two developments are specifically important: logicism: the attempt to reduce mathematics to logic and intuitionism: the attempt to develop mathematics from a purely constrictive perspective. It is interesting to note that that both attempts are reductive and in their pure form lead to the sacrifice of parts of mathematics as not well-founded.

The logicist program, that is associated with the work of Frege, Russell, Whitehead and the early Wittgenstein, could be seen as an attempt to deploy formal logic as a universal language for science. Wittgenstein's Tractatus can be interpreted as a philosophical analysis of the consequence of such a language. This research program was taken up by the members of the Vienna circle: with Carnap's 'Der logische Aufbau der Welt' as a central publication. The philosophical problems identified by Wittgenstein and the Vienna circle were soon overshadowed by technical problems in the heart of the logicist program itself (heuristics, incompleteness, undecidability, probability). Problems that are still in the center of current research. The modern notion of a universal language circles around the definition of the bit as a fundamental unit of information and recursive manipulation on binary strings as fundamental syntactic operation.

The metamathematical aspects of these operations are well-studied. At least part of Francis Bacon's original ambition is realized by modern information theory.

Brouwer's intuitionism deserves special attention in the context of a possible philosophy of information. In his first Kritik Kant had tried to give a foundation for the sciences, taking up the challenges that had been formulated by the empiricists. Kant's attempt took the form of an intricate analysis of the interplay between the act of thinking and the intuition (Anschauung) of space and time. In the center of this attempt we find an analysis of the transcendental unity of apperception in which the mind 'observes' the unity of its own actions. In this analysis Kant thought he had found the basis for a transcendental demarcation program, separating true scientific judgements from mere speculation. The influence of this research program in the 19th century is enormous. By the end of this century the program had ran in to serious difficulties caused by, amongst other things, the discovery of non-euclidian geometry. Several philosophers attempted to revive the Kantian program. In his 'Philosophie der Artimetik' Husserl for example tried to find a basis to concept of number in analysis of the act of counting. He abandoned this attempt after sharp criticism from Frege.

The predominant form Kantian philosophy at the end of the 19th century was based on a psychologistic (and in my view overly simplistic) interpretation of the human mind as an entity capable of observing its own actions (von Hartmann, Dèr Mouw, Heymans). This philosophical position is the starting point of Brouwer's foundational program: only the I and its experiences exist. Mathematics has to be reinterpreted as a construction of the human mind. Whatever we may think of this philosophical position, with the conceptualization of the Turing the conception of mathematics as a constructive activity proved to be very fruitful. The transformation from mathematics as a construction of the mind to mathematics as something that is computed on a machine is philosophically interesting. Since the middle of the 20th century the predominant view is that mathematics do not need separate philosophical foundation (Quine, Putnam), but the question of the adequacy of formal models of the world that we develop is still a central philosophical problem. The conception of a Turing machine can be seen as a modern distant relative of the Kantian transcendental unity of apperception. Just as Kant tried to find a foundation for science in the analysis of the act of thinking embedded space and time one could see the Turing machine as its materialistic transcendental counterpart. It is embedded in a conception of space and time and it is as close as we can get to a device for the analysis of universal models. Although I do not believe that this insight in itself leads to new fruitful philosophical theories it is an interesting historical parallel and certainly one that most computer scientists are not aware of. If anything, such reflections show that central issues in the philosophy of information are deeply rooted in the history of philosophy.

4 Philosophical Problems

In this paragraph I mention, without further analysis, some philosophical issues that I believe are of central importance in the philosophy of information.

Probability

It is impossible to write a philosophical analysis of the notion of information without a deeper study of the underlying problem of probability. The concept of empirical probability seems to come closest to what Shannon had in mind in his seminal paper, Shannon [1948]. It is mainly a theory about optimization of codes. However if we 'plug' the notion of probability rational degree of belief in to the formal framework of Shannon we get a formal theory about beliefs. If I believe that it is very improbable that life on earth generated form outer space then a statement from a reliable source implying that this is in fact highly probable contains a lot of information for me. Shannons formal theory is neutral and open to both interpretations. This dichotomy runs right through this book. Harremoës and Topsoe describe the traditional Shannon interpretation. The contributions of Dretschke and Seligman seem to be rooted in the rational belief position (although the former maintains that his theory is open to both objective and subjective interpretations of the concept of information) and Grüwald and Vitanyi present the notion of Kolmogorov complexity, that allows us to define the concept of an a priori probability of individual binary objects. Apparently this is a hairy issue. Hájek [2002] distinguishes a number of interpretations of probability: Classical Probability, Logical Probability, Frequency Interpretations, Propensity Interpretations and various modes of Subjective Probability. I am not convinced that all these notions of probability lead to fundamentally different concepts of information, but obviously a deeper analysis of the concept of probability is of vital importance for the development of a philosophy of information.

Meaning, information structure and randomness

Suppose that we reserve a room at the university of Amsterdam for the purpose of an 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:

1. A *random process* that generates bits (e.g. a dwarf flipping a coin).
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. 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 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. It contains an infinite amount of information that can never be learned in a finite amount of time.

Theorem 4.1 *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 is impossible to assign any form of objective probability to a source of information.

There is however a form of subjective probability that is very relevant in this context. 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 $1_1 1_2 \dots 1_n$. 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. 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 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 a set of phenomena, estimate the randomness deficiency and formulate a theory. Unfortunately the situation is more complicated. This becomes clear if we analyze the following claims.

Claim 4.2 *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 4.3 *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 4.4 *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.

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.

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.

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. 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.

The cooperative universe

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. In such a universe MDL therefore might be a viable methodological principle. It coincides with another well known dictum of Einstein: God is cunning, but he is not malicious. (!!! ref). 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.

Related philosophical problems

There are a number of issues that are tangential to problems in the philosophy of information. I do not consider them to be part of a philosophy of information per se, but they tend to be part of current philosophical debate. I mention four of them.

- Information and Virtual Reality. The idea that the mind can be fooled in to a purely subjective experience of reality without any objective substratum is is a recurring philosophical theme since Descartes. With the emergence of multimedia technology this possibility has gained new actuality. Some philosophers argue that with high probability we already live in a virtual reality (Nick Bostrup? !!! Ref).
- Discrete versus continuous models. If our universe is continuous it can

never be adequately modelled in terms of computational process. Quantum physics seem to imply that the the physical world is essentially discrete, but quantum mechanics itself is still object of fierce philosophical debate.

- Information and ethics. Being ethical seems to imply a certain obligation to give the right information. In this sense an understanding of the phenomenon of information has ethical implications. Some philosophers have tried to develop an ethics of virtual reality (Ref !!!)
- Information and esthetics: Our brain is an information processing device. There is evidence that our subjective experience of beauty can at least partly be explained in terms of information processing. The notion of idealization in art for instance can be interpreted in terms of two part code optimization.

5 Conclusion

It seems that the ambition to break codes, decipher messages, encode information, resolve conflicts between sources of information and combine data from different sources in a military context has been a driving force behind the emergence of the modern scientific notion of information. Another motivation for the scientific study of information was the development of new technologies for communication: telegraph, telephone, radar etc. These technologies confronted engineers with practical problems like noise, redundancy, multiple signals over one channel and efficient coding systems. Last but not least, the tremendous impact of the second world war on the development of the electronic computer for military purposes is well-known.

The modern notion of information is a creation of the joint effort of soldiers, spies, politicians, physicists, mathematicians, engineers and philosophers. The influence of mainstream philosophy on this process is limited. The impact of theory of information on philosophy is fundamental, not only in disciplines like methodology of science and theory of knowledge, but also in ethics and esthetics. Term 'information' has made a remarkable historical u-turn. It started as a technical term in antiquity, found its way to colloquial speech in modern times and then gained a completely different technical meaning in a military and political context at the end of the 19th century. With a new formal foundation developed in the 20th century it started to influence philosophical thought again as a fundamental discipline.

If we look at Kant's famous three questions: 'What can I know?' 'What must I do?' and 'What can I hope for?' then the first question at least partially involves a reflection on information. It is also immediately clear that a philosophy of information never can replace the whole of philosophy, since the concept of information gives us at best very limited assistance when pondering the second and third question. According to Kant philosophy should not reach for knowledge of the transcendent but it should be transcendental: i.e. it should study the necessary a priori conditions for the possibility of knowledge. In a surprising way modern information science opens a fascinating perspective on such a program for a methodological foundation of the sciences. This program might be somewhat removed from Kant's original ambitions, it is certainly transcendental in the sense that it allows us to formulate a priori conditions for the possibility of the growth of scientific knowledge with unprecedented mathematical precision and clarity. If one adopts 'the computational view' a number of new and powerful solutions to age old philosophical problems present themselves.

An example: if one lives in a world in which events are generated by computational processes than the Solomonoff-Levin distribution (or universal distribution) assigns an a priori probability to events. This probability is related to their computational complexity. This gives us an entirely new perspective on Hume's induction problem, the study of heuristic search and the analysis of human creativity and learning.

Another example: the concept of 'cognition as computation' allows us to formulate a partial a priori answer to the question of what can be known. In order for something to be knowable it must be computable. We have deep results on what can be computed and what not. We also know a good deal about the complexity issues involved.

A third example: modern logic studies epistemic logics, dynamic logics, non-well-founded set theory, update logics, belief revision systems and a myriad of related formal systems that give us an entirely new perspective on the questions of what can be known and what we could believe and how these questions are interrelated.

In this context a philosophy of information can be characterized along three dimensions:

- Foundational Nature: As soon as we think, reason or talk about reality, elements of information and computation are involved. Rationality, Computation and Information seem to be intertwined.

- Transcendental Nature: fundamental aspects of human knowledge can be brought under the rigor of mathematical proof, stipulating what is possible and what not.

- Fundamental Openness: there is no closed interpretation of the world. There is no universal heuristic method. There is no optimal heuristic method of Science, art and philosophy are open disciplines of potentially limitless complexity

Of course the question whether the 'computational paradigm' is correct is a matter of philosophical debate. This debate however can never be conducted properly without a thorough analysis of the philosophical issues involved. It would also be superficial to address these problems without a deeper understanding of the role of information and computation in various sciences like physics, mathematics, biology, linguistics and cognitive science. In each of these disciplines 'information' and 'computation' play a vital role, but prima facie in very different ways. There seems to be no direct route from the study of the abstract notion of information to practical results in these disciplines. The study of information has a definite empirical component as it takes a different guise in various parts of reality: physical, biological, social or psychological. At the same time progress in modern science depends critically on the use of computers to manipulate large quantities of data, to facilitate co-operative work between groups of scientists and to calculate the consequences of complex models of substructures of the world around us. This last observation gives a very practical motivation for a foundational study of 'the computational paradigm'. In an even broader perspective it is clear that the use of computers has a profound influence on our cultures and our societies. Computers change the way we communicate and the way we work. They affect our art and our science and ultimately the way we think about our self.

References

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

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

Clausius, R. (1850). Über die bewegende kraft der wärme und die gesetze welche sich daraus fr die wärmelehre selbst ableiten lassen. *Poggendorffs Annalen der Physik und Chemie*, 79:368–97.

(ed.), P. E. (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.

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

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.*

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

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.*

Shannon, C. E. (1948). A mathematical theory of communication (part i). *BSTJ*, 27:379–423.

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.