Version

1.0

THE Π RESEARCH NETWORK

Patrick Allo, Bert Baumgaertner, Simon D'Alfonso, Nir Fresco, Federico Gobbo, Carson Grubaugh, Andrew Iliadis, Phyllis Illari, Eric Kerr, Giuseppe Primiero, Federica Russo, Christoph Schulz, Mariarosaria Taddeo, Matteo Turilli, Orlin Vakarelov, Hector Zenil.

THE PHILOSOPHY OF INFORMATION

AN INTRODUCTION

The Philosophy of Information
An Introduction

Table of Contents

List of Figures

Preface

P

PREFACE

The Project

In April 2010, Bill Gates gave a talk at MIT in which he asked: 'are the brightest minds working on the most important problems?' Gates meant improving the lives of the poorest; improving education, health, and nutrition. We could easily add improving peaceful interactions, human rights, environmental conditions, living standards and so on. Philosophy of Information (PI) proponents think that Gates has a point – but this doesn't mean we should all give up philosophy. Philosophy can be part of this project, because philosophy understood as *conceptual design* forges and refines the new ideas, theories, and perspectives that we need to understand and address these important problems that press us so urgently. Of course, this naturally invites us to wonder *which* ideas, theories, and perspectives philosophers should be designing now.

In our global information society, many crucial challenges are linked to information and communication technologies: the constant search for novel solutions and improvements demands, in turn, changing conceptual resources to understand and cope with them. Rapid technological development now pervades communication, education, work, entertainment, industrial production and business, healthcare, social relations and armed conflicts. There is a rich mine of philosophical work to do on the new concepts created right here, right now.

Philosophy "done informationally" has been around a long time, but PI as a discipline is quite new. PI takes age-old philosophical debates and engages them with up-to-the minute conceptual issues generated by our ever-changing, information-laden world. This alters the philosophical debates, and makes them interesting to many more people – including many philosophically-minded people who aren't subscribing philosophers.

We, the authors, are young researchers who think of our work as part of PI, taking this engaged approach. We're excited by it and want to teach it. Students are excited by it and want to study it. Writing a traditional textbook takes a while, and PI is moving quickly. A traditional textbook doesn't seem like the right approach for the philosophy of the information age. So we got together to take a new approach, team-writing this electronic text to make it available more rapidly and openly.

Here, we introduce PI *now*. We cover core ideas, explaining how they relate both to traditional philosophy, and to the conceptual issues arising all over the place – such as in computer science, AI, natural and social sciences, as well as in popular culture. This is the first version, for 2013. Next year we'll tell you about PI 2014.

We hope you love PI as much as we do! If so, let us have your feedback, and come back in 2014. Maybe some of you will ultimately join us as researchers. Either way, enjoy it.

Yours, Patrick, Bert, Simon, Nir, Federico, Carson, Phyllis, Andrew, Eric, Giuseppe, Federica, Christoph, Mariarosaria, Matteo, Orlin, and Hector.

CONTRIBUTORS

With thanks to…

Although the whole book has been a cooperative exercise, with authors reading and commenting on each other's chapters, and frequently contributing material to each other, we acknowledge the primary sources of the material of the chapters here. Particular thanks are due to:

1	A quick history of the philosophy of information	Andrew Iliadis
2	What is the philosophy of information now?	Phyllis Illari
3	Naturalised information	Christoph Schulz
4	Ethics	Matteo Turilli
5	Society	Mariarosaria Taddeo
6	Meaning	Eric Kerr
7	Truth	Simon D'Alfonso
8	Knowledge	Simon D'Alfonso
9	Science	Federica Russo and Phyllis Illari
10	Cognition	Orlin Vakarelov
11	Mind	Bert Baumgaertner
12	Logic	Patrick Allo
13	Computation	Giuseppe Primiero and Nir Fresco
14	Algorithmic information theory	Hector Zenil
15	Personal identity	Phyllis Illari and Federico Gobbo
	Illustrations	Carson Grubach and Orlin Vakarelov

This book has been supported by the Bass Connections Initiative and Information Initiative at Duke University, as well as by a New Faculty Fellows award from the American Council of Learned Societies, funded by the Andrew W. Mellon Foundation. We are very grateful for the support.

Part I: Introductory material

Chapter 1

1. A QUICK HISTORY OF THE PHILOSOPHY OF INFORMATION

Mapping many thinkers (Beta Chapter)

1.1 Introduction

> *'It is hardly to be expected that a single concept of information would satisfactorily account for the numerous possible applications of this general field.' (Shannon, 1993, p. 180)*

To those who work with mathematical concepts of information every day, the answer to the question "why worry about information?" will be blindingly obvious. Even those unfamiliar with the mathematical concepts may be able to see for themselves that information has come to be very important to us, but remain bewildered as to where this explosion in talk of information came from.

We could begin in many different places, but choose to open with the vision of the brilliant British mathematician, philosopher and engineer Charles Babbage (1791-1871), who invented a peculiar machine known as the "difference engine", a feat of engineering that can easily be seen as a precursor to the modern computer, although it did not actually run any programs. The difference engine was essentially a calculating machine, automatic and constructed in order to erase the errors in calculation that were so common in Babbage's time. (Previously, most calculating was done in the head, or aided by pen and paper.)

The machine that Babbage envisioned would have been a massive undertaking, consisting of an incredibly large number of wheels, vast quantities of brass, and many different cogs and sector gears. The government at the time had spent a lot of money financing Babbage's quest to build the machine. Sadly, the difference engine was never actually built during Babbage's lifetime. However, after a brief period in the shadows, the idea of the machine became popular once again and this led eventually to the successful real-world construction of a difference engine. The importance of Babbage's contribution is that he took the theory of numbers out of the world of abstraction and into the realm of technology, computability, and engineering. Babbage is the first figure in our history of the philosophy of information who thought about bringing the abstract notion of calculation into the materiality of life.

In this chapter we will follow the progress of the idea of information into philosophy, via technologies now in daily use. We explain the beginning of the mathematical approach to information, and how it enabled the creation of concepts, followed by technology and further concepts, that revolutionised our world. There are introductions to the profound thinkers who should be much more widely understood than they are –

particularly Alan Turing (1912-1954) and Claude Shannon (1916-2001) – but these introductions can be very formal and mathematical. Here, the key aim is to convey how the insights of these mathematicians came to have such profound philosophical implications. We will look first at Turing, then at Shannon, before turning to an examination of how their ideas generated novel ideas applicable in many different places.

Some of these places include exciting schools of philosophical thought all around the world. We will look at Norbert Wiener and the "Cybernetic" school, which focused on how to use information to control nature and regulate ecology. Then we will examine the French continental school, beginning with Gilbert Simondon, who sought to articulate new informational concepts using their own brand of speculative realism. Finally, there are towering figures such as Fred Dretske, (1932-2013) who contributed so much to the philosophy of information debates in epistemology and knowledge. We will look at all of these figures, and more, in this chapter.

1.2 Turing's basic idea

What did Alan Turing give us? Why might he be considered the father of modern computer science? Turing gave us a lot of things, and it is easy to get distracted trying to understand the details of his work. But one way of understanding Turing's vital importance is by seeing him as the person who gave us the idea of *computation*. This is the idea that we could take many different procedures that at first sight look quite different, and reduce all of them to different sets of basic operations that could be run by a machine, separating the operations themselves from the brain normally used to run them. The other side of this idea was that of a machine that was not a coffee-maker, washing machine, or lawnmower, but a machine that could run these many different procedures by running these basic operations on data to produce more data – a kind of information processing.

Figure 1: Alan Mathison Turing (1912–1954) © National Portrait Gallery, London

The machine described is a computer, of course. What makes your computer, and also your smart phone and your iPad, special among all the machines you have is that they can do a lot of different things (your other machines can only do one or two things). Looking back, it can be difficult to grasp how startling this idea was: a machine created not to do some specific thing, like make coffee, wash clothes, or mow grass, but to do indefinitely many things, depending only on the ingenuity of the programs created for it.

Turing's main way of giving us these ideas is by way of what came to be called a "Turing Machine" (Turing, 1936). Turing was concerned with what can and cannot be calculated or computed and how. He

came up with the idea of a machine that mimics a human calculating using pen and paper (see figure 2). Turing suggested that the operations of the human comput*o*r (to distinguish it from an artificial comput*e*r) performing some task may be completely mechanized by breaking down the rules of computation into a series of basic sub-rules. We can consider the human computor making calculations easier by using her notebook, and breaking down her operations page by page. Imagine her following instructions to alter the page she is on, or turn to another page. Simplify by thinking of the content of each page as replaceable in principle by a single symbol, and think of a very long notebook, so the human computor never reaches the last page. We have something like an infinitely running paper tape, with symbols on page after page after page. The human computor moves from page to page, following the instructions either to change the symbol or to move, according to the symbols. She could use this procedure to do many things, such as add numbers.

Figure 2: Computors in an accounting office (1924) (Source: Library of Congress)

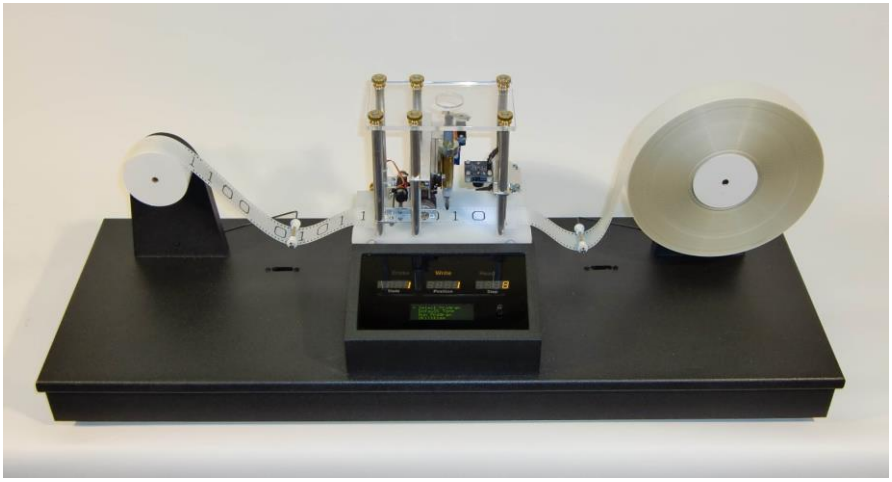

Figure 3: A modern reproduction of a Turing Machine. (Courtesy of Mike Davey)

Turing argued that either human computors or some mechanical devices could perform computation, when this is understood as finite sequences of such operations on symbols – operations either to change the symbol or to move – as described above (see figure 3). Remember that Turing's contribution wasn't

an actual machine – it was *the idea of this possible machine*. The Turing machine was an abstract, idealized representation of a human computor, whose operation is determined by discrete, effective steps: at each step it is entirely defined what the computor is allowed to do. This machine was a hypothetical mechanical device with unlimited storage capacity (an infinite tape to write on, like the infinite sequence of pages in the notebook) and a limited set of possible actions (defined by its table of instructions). It was proven that such a general and simple idea would be computationally equivalent to almost *any* conceivable digital computing system. In this way, Turing defined the computation process itself. And once this is defined, it doesn't matter whether it is done by a person with a notebook, or by a machine working from a hard disc or USB drive or – in Turing's mind – a tape and a scanner.

In this case, we imagine a human computor performing some particular task, and understand that if a machine follows the same instructions as the person, the machine will perform the same task. So far this is describing a machine something like a simple calculator, able to perform a limited range of tasks, but nothing like a modern computer. But Turing generalized his idea further to the universal Turing machine, by showing that it is possible to construct a single universal machine that can be used to compute any function that is computable on a standard special-purpose machine (Turing, 1936, pp. 241-242). If the universal machine is supplied with a tape, at the beginning of which is the table of instructions for some special-purpose computing machine, then the universal machine will compute the same function as the special-purpose machine. But so long as different instructions can be loaded onto the universal machine, it will be able to compute multiple other functions, of many other special-purpose machines. The exact construction of the universal machine doesn't matter here. For our purposes, think of the operation of the universal Turing machine as explained by its execution of the instructions of some special-purpose Turing machine. Now, the universal Turing machine is like a modern computer, which can upload various different software programs to enable it to do indefinitely many different kinds of computations. This is how Turing gave us the idea of a computer. (See section 13.3b for further discussion.)

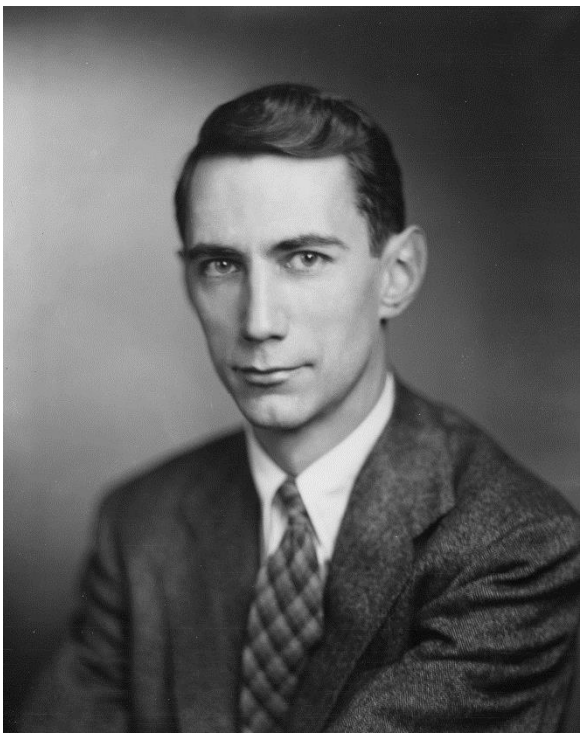

Figure 4: Claude Elwood Shannon (1916-2001) (Courtesy of MIT Museum, Boston / Nixdorf Museums Forum, Paderborn)

1.3 Shannon's basic idea

What did Claude Shannon give us? He created a new branch of maths known as mathematical communication theory. That was very important to maths, but why does it matter outside maths? At heart, Shannon gave us a very general language in which to describe precisely many very different things. This is the language of information. The connection with Turing's work is immediate once you realise that the inputs and outputs of a Turing machine are information. Turing gave us the *idea* of a computer, which processes information. The *idea* of information already existed. But Shannon gave us the language in which we can describe the bits required for a computer programme and the bandwidth of an internet connection, and all the things that are essential to make computers actually work. Shannon allowed us to *quantify* information – say how much information there is. As computers and informational technology like the internet have become so much a part of

our daily lives, this language of information (now there are multiple languages) has become increasingly important to navigate the world successfully. Further, we will see in the next section how the idea of a general language of information, now that we have it, can also describe many more things.

Shannon showed in 1948 how information could be transmitted efficiently across communication channels using coded messages. So Shannon described an information-generating system as a combination of five essential components: an information source, a transmitter, a channel, a receiver and a destination (Shannon, 1948, p. 4; Wiener, 1948, p. 79). The information source produces a message to be communicated to the receiver. The transmitter operates on the message to produce a signal suitable for transmission over the channel, which is simply the medium of signal transmission. The receiver reconstructs the message from the signal. Finally, the destination is the entity for which the message is intended. According to Shannon, communication amounts to the source of information producing a sequence of symbols, which is then reproduced by the receiver. The reproduction is only to some degree of accuracy – as we find when unable to distinguish every word clearly during a mobile phone conversation.

Taking this description of communication, Shannon attempted to solve the 'fundamental problem of communication' (Shannon, 1948, p. 1), finding the optimal way to reproduce, exactly or approximately, messages at their destination from some source of information. One vital but easily missed novelty in Shannon's work is that his information theory has abstracted away from the physical media of communication, so that relevant physical constraints can be analysed separately. It doesn't matter whether you are using a phone – mobile or landline, or what make or model – or any of many different radios, or a Skype call, or sending a large file by email. Shannon's theory is about the *information transmitted* and tells you about that information, whatever the particular physical medium you are using to transmit it. Shannon provided a statistical definition of information as well as general theorems about the theoretical lower bounds of bit rates and the capacity of information flow – which tell you how fast you can transmit information.

Importantly, on Shannon's analysis of information (there are now many other analyses), information does not involve any meaning. It doesn't matter whether you are trying to send "Meet you for lunch on Tuesday 2pm", or "8459264628399583478324724448283". Shannon information concerns only correlations between messages, variables, etc. For example, it concerns whether the message received matches – correlates with – the message sent. It is a quantitative measure of how much information is successfully conveyed. Shannon information is much more specific than our ordinary usage of "information"; Shannon information tells us nothing about whether a message is useful or interesting. The basic aim is coding messages (perhaps into binary codes like 000010100111010) at the bare minimum of bits we must send to get the message across. One of the simplest unitary forms of Shannon information is the recording of a choice between two equally probable basic alternatives, such as "On" or "Off". A sufficient condition for a physical system to be deemed a sender or receiver of information is the production of a sequence of symbols in a probabilistic manner (Wiener, 1948, p. 75).

Shannon's mathematical theory is still used today in "information theory", which is the branch of study that deals with quantitative measures of information. Two of Shannon's metrics are still commonly used: one a measure of how much information can be stored in a symbol system, and the other a measure of the uncertainty of a piece of information. The English anthropologist Gregory Bateson famously defined information as "a difference that makes a difference." This definition aptly characterizes Shannon's first metric. Binary is the code usually used by computers, representing everything using only two symbols, 0 and 1. One binary digit, or bit, can store two pieces of information, since it can represent two different

states: 0 or 1. However, two bits can store four states: 00, 01, 10 and 11. Three bits can store eight states, four sixteen and so on. This can be generalized by the formula log 2 (x), where x represents the number of possible symbols in the system. Log 2 (8), for instance, equals 3, indicating that three binary bits of information are needed to encode eight information states. These possibilities are essential to how much disc space is needed to store a computer program.

Shannon's second metric is "entropy," a term recommended to him by John von Neumann because of its relation to entropy in thermodynamic systems. Some say this use of the term is fortunate because it captures similar phenomena, but others say it is unfortunate, because the two types of entropy are only related to a certain extent – they are only somewhat isomorphic. Simply put, entropy in thermodynamics measures disorder. But information entropy is a measure of uncertainty in terms of the unpredictability of a piece of information. Information that is highly probable hence more predictable has a lower entropy value than less distributed information, and therefore tells us less about the world. One example is that of a coin toss. On the one hand, the toss of a fair coin that may land heads or tails with equal probability has a less predictable outcome, higher entropy, and thus a greater ability to decrease our ignorance about a future state of affairs. On the other hand, a weighted coin that is very likely to fall "Heads" has a very predictable outcome, lower entropy, and therefore is unable to tell us anything we do not already know.

The significant aspect of Shannon information is that a produced message is selected from a set of possible messages. The more possible messages a recipient could have otherwise received, the more surprised the recipient is when it gets that particular message. Receiving a message changes the recipient's circumstance from not knowing something to knowing what it is. The average amount of data deficit (uncertainty or surprise) of the recipient is also known as informational entropy (Floridi, 2011c, Chapter 3). The higher the probability of a message to be selected, the lower the amount of Shannon information associated with it is.[1]

The importance of uncertainty to Shannon information is significant. Although Shannon is not concerned with what messages *mean* to us, the amount of Shannon information conveyed is as much a property of our own knowledge as anything in the message. If we send the same message twice every time (a message and its copy), the information in the two messages is not the sum of that in each. The information only comes from the first message, while the second message is redundant. Still, for Shannon the semantic aspects of messages carrying meaning are 'irrelevant to the engineering problem [of communication]' (Shannon, 1948, p. 1). This, then, is how Shannon showed us what could be done with the concept of information.

1.4 Extension of the concepts

From Turing and Shannon we get the idea of an artificial machine that can process information, and the basic language of information necessary to build such a machine – to build a computer. The creation of computers and everything that has followed is a dramatic change in the world. But the profound importance of the language of information is that it has altered the way we think forever.

The key to this alteration is in the abstraction away from the peculiarities of a particular unique physical object in front of you. Turing began the separation between the unique and particular machine before you (your shiny red laptop, with the dent on the case from where it nearly departed this world in a freak skiing accident, with the photos on it from that skiing holiday), and the multiple other machines very like it, and the many processes they can perform. Shannon pushed that even further by giving us the language to

[1] Shannon information is defined as the base two logarithm of the probability of selecting a message *s*: $\text{Info}(s) = \text{Log}_2(\text{Prob}^{-1}(s))$.

describe the information being processed, independently of what is doing the processing. The information revolution has been thriving on this separation ever since.

Once these refined concepts of information were established, they were applied to many things. The idea of information, and the multiple mathematical formalisms now available to quantify information, proved enormously fruitful in creating shared conceptual schemes that allowed our investigations to leap forward in so many different fields. To this day, many thinkers from disciplines far removed from mathematics, science and engineering find use for the notion of information. Psychologists, sociologists, anthropologists, and historians have each turned to the notion of information to see what it has to offer their own disciplines, and often this occurred through the curious probing of philosophers. Today, topics as diverse as biopolitics (which studies things like the power of statistics to control human populations over time) and ontology (the philosophical study of being) gain something by considering the many facets of the notion of information. It is this wide applicability of these notions that meant that something that started as maths became relevant to multiple scientific fields, helped us build a new world, and began shaping our society, our ethics, how we understand knowledge, and everything else.

We will now introduce two groups that greatly influenced what the philosophy of information would become at the turn of the twenty-first century, leaving current philosophers of information for the rest of this book. The first of these groups remained closely tied to the original insights of Turing and Shannon, slowly adding to them bit by bit (no pun intended), while the second, it could be said, sought instead to ask what the notion of information might mean ontologically in terms of more "macro" discourses (economics, politics, sociology, etc.). While some of these thinkers were not mathematicians or scientists in the proper sense, they knew enough about the theoretical significance of Turing and Shannon's ground-breaking work to be able to think through its consequences. While this might be oversimplifying a bit, these two groups remained related, producing thinkers who sought to ponder the consequences of information after the initial contributions from Turing and Shannon in the twentieth century. Though some were less concerned with mathematical proofs and empirical observations than others, each sought to consider the implications of information for realms beyond those of numbers and theorems proper.

1.5 Cybernetics

The first group was a loose-knit collective of American scientists, communication theorists, and psychologists that focused on work that, for the most part, came to be grouped under the label of "cybernetics". Cybernetics, simply put, became a way to talk about systems and structures of information as applied to "larger", "closed" ecological fields, such as biology, society, nature, psychology, or economics, rather than in terms of, say, a rudimentary telephone system. In many ways, cybernetics was synonymous with control. It dealt with constraints, and in so doing attempted to demarcate systems whose channels of feedback (both "positive" and "negative") amounted to the proper or improper functioning of the overall system. A brilliant American mathematician named Norbert Wiener (1894-1964) was largely responsible for the initiation of this field, writing a book with the word "cybernetics" in the title (Wiener, 1948), in large part because he considered the application of Shannon's theories to the fields of biology, ecology, and society.[2] The notion of information feedback in nature was one of Wiener's favourite topics. The reason for this was that he was interested in seeing the "measure of reality," as it were. He was interested in its constraints and the degree by which nature might be controlled.

[2] Another important figure is Jay Wright Forrester, known as the founder of systems dynamics, but for our purposes he will not accompany us on this journey.

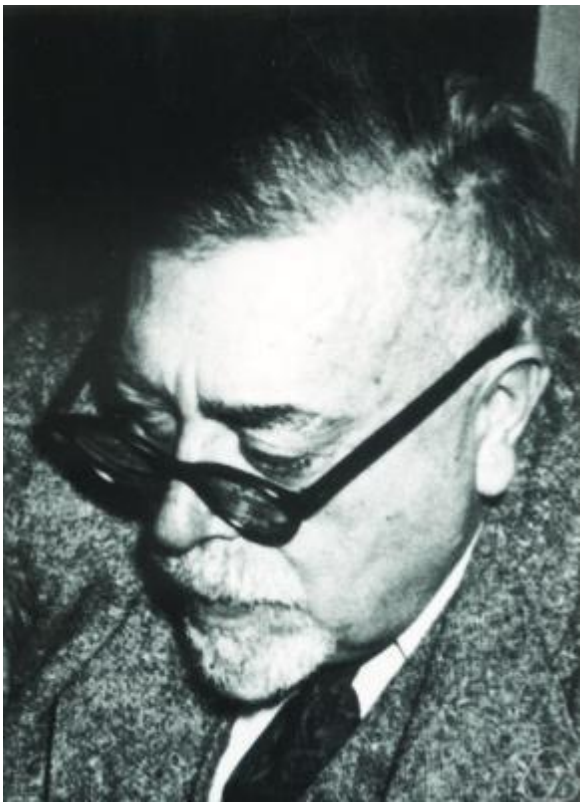

Figure 5: Norbert Wiener (1894–1964) (Courtesy of Konrad Jacobs)

Other individuals who sought to use the concept of information in a cybernetic sense included the psychologist Charles E. Osgood (1916-1991), the scientist and mathematician Warren Weaver (1894-1978), the communications scholar Wilbur Schramm (1907-1987), and, perhaps most importantly, the mathematician John von Neumann (1903-1957).

Von Neumann can in many ways be considered the spiritual guide of this group. Without diving into a detailed history (many already exist, and the reader is encouraged to review them, particularly (Dyson, 2012; Gleick, 2011)), von Neumann oversaw the development of a machine that would eventually become the physical realisation of Turing's thesis at the Institute for Advanced Study at Princeton. von Neumann (1958) was a speculative book, which unfolded a comparison of information and computation in the area of human biology, specifically the brain. In this book, which initially stemmed from notes for the Silliman Lectures that were to be delivered at Yale in 1956, von Neumann outlined his theory of the brain by comparing it, piece by piece, bit by bit, with the schematics of a digital computer. Before this text, no one had explicitly compared the functions of the human brain to functions associated with the computer. Topics such as memory, sense, and processing in human biology were interpreted as being analogous to computational functions (for more on cognition, see Chapter 10; for more on computation, see Chapter 13). Although von Neumann is not a household name, the comparison between the natural brain and the digital computer is intuitively appealing, and today, many people understand the analogy that refers to the brain as a type of "hardware", which can be seen as running a type of "software" called the mind, and this insight (although it is now controversial) flows from the comparison made by von Neumann.

Although von Neumann was centrally important to the development of information as a worldly phenomenon, Wiener became, in some sense, the next guiding light that championed the theoretical complexity of information beyond the sender-receiver model illustrated by Shannon. His contribution largely came in the form of a book (Wiener, 1948), in which Wiener sought to introduce the notion of information to the study of everyday encounters. Wiener saw communication as information just as Shannon did, yet where Shannon stated that he attempted to explain only an engineering approach to information and communication theory in his paper of 1948, Wiener sought a way that would allow Shannon's mathematical theory of information to lay the groundwork for a much more fluid and diverse conception of communication, developed from these connective underpinnings. Wiener wanted to explain how information made the world "tick".

The most interesting figure among the group of cyberneticists (the famous Cambridge philosopher Bertrand Russell had a few not-so-nice things to say about him), Wiener articulated further that cybernetics should seek to find the difference between information as an entity that can be sent and

received (the "transmission" model) from one that is semantically constituted in the flux of interpersonal communication. Here he predates Dretske, while also arguing for a non-probabilistic approach to semantic information. In a book originally published in 1950, Wiener wrote:

> It seems necessary to make some sort of distinction between information taken brutally and bluntly, and that sort of information on which we as human beings can act effectively or, *mutatis mutandis*, on which the machine can act effectively. In my opinion, the central distinction and difficulty here arises from the fact that it is not the quantity of information sent that is important for action, but rather the quantity of information which can penetrate into a communication and storage apparatus sufficiently to serve as the trigger for action
> (Wiener, 1988, p. 93).

Wiener developed an approach that was philosophically distinct from that of Shannon, one that articulated a world where semantic information remained different from, yet still tied to, traditional notions of communication, where the amount of data sent mattered perhaps less than the type of data that could 'penetrate into a communication and storage apparatus sufficiently to serve as the trigger for action'. (Wiener, 1988, p. 93) He helped to establish that there are different types of information. Diverse styles of informing mattered to the cyberneticists, as any careful reading of their work will show. Current debates on everything from psychiatry to philosophy of mind remain deeply tied to this distinction in terms of information, yet many, it would seem, are unable to account for the interplay between what Wiener called "brutal" or "blunt" information and the 'sort of information on which we as human beings can act effectively'. (Wiener, 1988, p. 93). This might be one of the most important problems that the philosophy of information seeks to uncover. Current philosophers such as Floridi (who we will introduce in Chapter 2) are attempting a systematic philosophy that might define the interaction between these two levels of information (and many more). Indeed, the emergence of philosophy of information as a recognised field is long overdue.

The tradition of separating these different types of information was extended by Osgood. He was an American psychologist whose work lay close to cybernetic concerns, and who is most famous for developing the connotative meaning of concepts known as the "semantic differential". Osgood (1952) acknowledged that there was a field beyond the strictly informational-theoretic terms developed in the area of engineering such as "sending" and "receiving", particularly in his description of "choice-parts". He intended these to be moments where the information-theoretic content of a message (in the transmission-probabilistic model) gives way to something not entirely predictable. This would be a theme throughout Osgood's career. He saw communication sequences as informational in the engineering sense, but also as something that might bring meaning in terms that are not directly related to the sharing of quantitative information, even in Dretske's sense of probabilistic semantic information. In 1952, Osgood said that "choice-points" were 'points where the next skill sequence is not highly predictable from the objective communicative product itself' (Osgood, 1952, p. 197.). To explicate this, he uses the simple example of having to explain that it is better to wait to wash a car. He writes: 'The dependence of "I'd better not wash the car" upon "looks like rain today," the *content,* of the message, reflects determinants within the semantic system which effectively "load" the transitional probabilities at these choice-points'. Osgood would go on to describe a theory that lay beyond the "predicative" model; however, this remained strongly tied to the transmission model of communication. Like the theorists of cybernetics, he theorized the way a semantic notion of information might be predicated on a strictly engineering perspective of communication, yet he reserved space for a non-connective realm.

The idea that the transmission model undergirds semantics and other modes of information and communication techniques makes sense given the utility of its wartime origins. Developed in the Bell Labs in New York City during the Second World War, its inventor and one of the individuals with which we began this chapter, Claude Shannon, was a brilliant young thinker who spent the better part of his academic life at MIT. His Master's thesis on Boolean algebra and what he called a "logic machine" (Shannon, 1938, p. 471) would lay the foundations for the design of computer circuits. One of Shannon's oft-quoted passages is the following, taken from his landmark paper:

> The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have meaning; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one selected from a set of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design.
> (Shannon, 1948, p. 1.)

This distinction between what we can call "data" and "semantic information" would be explicated by cyberneticists and others, including Weaver, Wiener, Osgood, and Schramm, each of whom believed that communication is, first and foremost, the flow of information. Clearly, the idea for Shannon is that the transmission approach does not have much to do with semantic information. Osgood and Wiener were equally vocal about the transmission model's inability to account for semantic information. The idea was not that the transmission model had *nothing* to do with semantics, but rather that, while it might *underpin* semantics, it cannot account for it on its own. The absence of this important distinction is unfortunately reproduced in general discussions that feed the popular imagination of what information theory and cybernetics is all about. Any number of cybernetic texts speak to the open place left within information theory that would later be taken up by many philosophers, some of whom will now be introduced.

1.6 Dretske

The philosopher Fred Dretske was a central figure in the (re-)establishment of interest in the concept of information in current Anglo-American analytic philosophy. He was not the first analytic philosopher to offer a systematic treatment of information. Prior to Drestske, most prominently, in the 1950s, Rudolf Carnap (1891–1970), together with the Israeli mathematician and linguist Yehoshua Bar-Hillel (1915–1975), attempted to develop an account of semantic information based on Shannon's theory. Much like the rest of Carnap's later technical work on formal semantics and inductive reasoning, the work on information was mostly ignored by his contemporaries. Mainstream analytic philosophy had turned away from formal languages and towards "ordinary language philosophy". Dretske's work on information, developed most fully in his book *Knowledge and the Flow of Information* (KFI) (Dretske, 1981), came at a time when the tides of analytic philosophy were changing. There was renewed interest in formal modelling from scholars working in areas such as epistemology, the philosophy of mind and the philosophy of language. Dretske's work was not in philosophy of information proper, or at least this is how he saw it. Even in his last days, he was expressing scepticism about the PI project – a view that, we believe, was

reversible.[3] His was a project in epistemology. As he described it (paraphrasing): "I was trying to develop an externalist theory of knowledge. I thought that what the engineers were doing with the concept of information may be very useful for philosophers and for the problem of knowledge."[4] What allowed Dretske to be successful in bringing (back) the concept of information to philosophy was the fact that he managed to link it to the way philosophers (at the time) discussed other problems. He did this through a skilful use of intuitive examples intertwined with formal theory. He was, indeed, a master of this. The success was, however, only partial. The epistemological theory that he suggested was quite controversial, mostly because externalism fell on hard times. In the mind of many philosophers, Dretske's attempt to use information to understand knowledge was seen as more or less a failure – information was tried, it did not work, so we should not bother. Dretske's own work moved in the direction of philosophy of mind. While the ideas from KFI were clearly under the surface of his work, they did not recur prominently. It was not until the epistemological work of Floridi twenty years later that the concept of information re-entered the discussion of the same problems – in a quite different and more fundamental way.

Now let us look more closely at the key ideas of Dretske. They will be developed more fully in many of the other chapters of this book. The driving idea behind the connection between information and knowledge for Dretske is really rather simple. When we want knowledge, what it is that we really want? What is it that we value? What are we willing to pay money for? The answer for Dretske was: we want information. Information is the *prime epistemic resource*. This is not truth, or justification, or something that reliably causes your beliefs about the world, or to be infallible, or to know that you know. Many of these are important, but they are secondary or implicit. We go to a train station and want to know at what time the train departs. We need to find information about this, so we go to the information booth. This (or other similar places) is where the valuable resource can be gotten.

Dretske held the view, known as the strongly semantic view of information (see Chapter 7), that one cannot have information if it is not true. It wouldn't be an epistemic resource otherwise. When we ask at the information booth for the train time we do not expect a statement on the topic of train times. We expect the truth about the train. Anything else means incompetence. Dretske's famous expression captures this idea: 'False information and mis-information are not kinds of information—any more than decoy ducks and rubber ducks are kinds of ducks.' (Dretske, 1981, p. 55).

Dretske never even attempted a theory of justification. His view was that, first, it was not necessary for (an externalist conception of) knowledge, at least for basic perceptual knowledge – the main goal of his theory. Second, a proper theory of justification would be too messy. About the question of knowledge of knowledge, his view was that only philosophers are interested in such questions. Ordinary people only want knowledge – they want information.[5]

How, then, can such an idea of information as an epistemic resource be made precise? First it must be clear that this conception of information must be semantic, as knowledge is a semantic phenomenon. Second, being an account of knowledge, the conception of information must provide a means for

[3] Interestingly, the very last undergraduate lecture that he delivered on April 17, 2013, as a guest in one of the authors' epistemology classes, was precisely on his informational account of knowledge. He was as enthusiastic about it as ever. Prior to the class, over lunch, he remarked that this was the very first time he had ever taught this material to undergraduate students.

[4] Here is a more direct quote: "Maybe – or so we may hope – communication engineers can help philosophers with questions raised by Descartes and Kant. That is one of the motives behind information-based theories of knowledge." (Dretske, 2008)

[5] This was important because for many philosophers the question of justification and epistemic reflexivity were central and decisive aspects of knowledge. The unwillingness, or inability, of Dretske to deal with them made his information approach appear too weak to resolve the philosophical problems of knowledge.

understanding how one may have beliefs. (Note that Dretske is dealing with the standard conception of knowledge that assumed that the vehicles of knowledge are beliefs. This is the so-called "doxastic" approach to knowledge. We will see later, in Chapter 8, that some philosophers, most prominently Floridi, adopt an informational approach precisely to avoid a doxastic approach.) Dretske thinks that there are enough resources in Shannon's communication theory to capture the needed ideas. His strategy is the following: (1) View the problem of *A* knowing that *s* is *F* (*A* having a belief of propositional content *F(s)* that meets some conditions of being an instance of knowledge) as a problem of there being a special kind of communication channel between a source and a receiver of information (a known and a knower). Knowledge is obtained by the receiver getting a special kind of message *m* from the source. (2) Provide the needed condition on the communication channel to allow a message to have the needed propositional content for the receiver. Beliefs are intentional states, yet the condition of the communication channel cannot offer the correct kind of intentionality for mental states needed for knowledge. Thus, (3) provide the needed further condition for intentionality for beliefs. Once this is done, knowledge can be defined as the agent having a belief that *s* is *F* generated (or supported) by the correct kind of information that *s* is *F*.

Step (1) is primarily a way of reconceptualising the problem of knowledge as a problem about information in the Shannon sense. The real work is related to step (2) and step (3). The details will be developed in later chapters. A proposition of the form *s* is *F* has an "object" (what it is about), and "properties" (what is ascribed to the object). So, *s* is the object and *F* is the property. In a communication channel, the object is fixed. It is the source of the information. The message we get from the source is about which of the possible properties (or states) the source may be in. The problem of Shannon's theory is that messages give us only a probability of the source being in one of the possible states. For Dretske, what is missing from the Shannon story is not the components needed to provide meaning (the object and the qualities), but the under-specification of the meaning. The solution is natural. Given some background information *k*, "a [message m] carries the information that *s* is *F* = The conditional probability of *s* being *F,* given *m* (and k) is 1 (but given *k* alone is less than 1)." This is called the "information content" of a message. In other words, a message has information content if it uniquely determines the state of the source. Note that not all messages have information content.

Such a conception of content has a well-known problem. If a message *m* has the content *s* is *F*, and *s* being *F* implies the *s* also has the property *G*, then *m* has the content *s* is *G*. This is because if the probability of *s* being *F* given the message is 1, then the probability of *s* being *G* will be 1 as well. Here is the problem: one may know that *s* is a square but one may not know that *s* is a rectangle. One may not even know what a rectangle is, as many children do not. If all that was needed for knowledge were to have information with some content, then children will not only have to know about rectangles as soon as they learn about squares, but they will automatically have to know all of mathematics. This difference is related to the issue of intentionality. Dretske's solution is to consider *how* information is carried by various states to the mind. He makes a distinction between a state's carrying information in digital and in analogue form. A message carrying information of content *s* is *F* is in *digital form* if it does not carry any further information content than *s* is *F*. If it carries more, it carries it in analogue form. With this, Dretske can separate the information content *s* is *F* from *s* is *G*, because only one of them may be carried in digital form. The message "*s* is a square" carries the information about squareness in digital form, while it carries information about rectangleness in analogue form. He thus defines another kind of content – "semantic content". This is the content carried in digital form. Only some information carriers carry information in

digital form. Dretske then uses this idea to provide an account of beliefs. The proper notion of content for beliefs is not information content, but semantic content. (See Chapter 3 for more on Dretske.)

1.7 French Philosophy of Information

The American cyberneticists were only one group born from the theories of Turing and Shannon. Another, more distanced group, was a closely-related team of predominantly French philosophers. They were responsible for expanding certain themes as well as the philosophical consequences of Turing and Shannon's ideas in the very early stages. The first of these thinkers were Raymond Ruyer (1902-1987) and Gilbert Simondon (1924-1989), who published mainly in the 1940s and 1950s, closely followed by Gilles Deleuze (1925-1995), Michel Foucault (1926-1984), and Jean Baudrillard (1929-2007), in the latter half of the twentieth century.

This second group practices something that is often popularly referred to as "continental philosophy" – a philosophy somewhat different from others that will be presented in this book. However, this distinction is quickly falling by the wayside in current debates. One key characteristic that tends to remain true of continental philosophy, however, is that it seems to focus more on the invention of concepts and terminology, which are expected to be used in future debates on whatever issue is at hand. The second characteristic is that continental philosophy tends to have a predilection for articulating problems and attempting to find the theoretical missteps in preceding theories, rather than in the simple prolongation of already established theoretical customs. Though this approach has been occasionally derided as contrarian, today many see the value of its aims and methods. While continental-type concepts such as "hyperreality" (the inability to distinguish between reality and a simulation) and "immanence" (the absence of hierarchy) might have appeared far-fetched to more mainstream branches of philosophy in the twentieth century, today topics such as these are no longer viewed as the domain of continental philosophers alone. Indeed, many continental theories appear to be completely compatible with the more analytic notions inherent in the philosophy of information.

While the current approach to the philosophy of information has begun by analysing the texts of philosophers whose work relied heavily on philosophy of mind and epistemology – perhaps most importantly the work of Dretske – the French philosopher Gilbert Simondon remains a key figure in the history of the articulation of a robust philosophy of information. As we have seen, the American cyberneticists knew that there were areas yet unexplored by the concept of information as expressed in semantics. Simondon knew this as well, and his approach to information was, in a way, an extension of these concerns. While he remained deeply critical of some of the cyberneticist approaches to information, he did not disagree with the engineering notion of information entirely. Like some contemporary philosophers of information, he sought to push the notion of information to an even more "naturalistic" extreme.

Simondon's approach to what we can call "informational ontology" is a type of phenomenological (subjective, observable) extension of the mathematical theory of communication, but also one that accounts for the indeterminacy (openness, unpredictability) of information's interactive existence in terms of biological and technical structures, thus furthering the concerns of the earlier cyberneticists. Simondon approached information from a perspective that allowed for the *interoperability* of different types of information, leaving space for an *indeterminacy* that would remain a fundamental component of Simondon's open informational schema. Where the cyberneticists argued for control, Simondon argued against the automation of phenomena. A simple way to think of the difference is this: where the

cyberneticists thought about information as a bunch of "closed" systems in the world, Simondon's approach held that there are no systems "as such", but only the differences that information introduces in a universe that can be thought of as one giant, single system of information in action.

These two factors – interoperability and indeterminacy – would allow Simondon to apply the notion of information to fields beyond pure mathematics, and to mix and match various forms of information in order to come to some understanding of informational and technological genesis. He sought, for example, to think of the evolution and genesis (change) of objects, technology, and the world, in terms of information. Indeed, his two main theses (Simondon, 1958, 1964), originally defended in 1958, are chockfull of metaphysical ideas pertaining to a naturalized understanding of informational genesis.

Another simple way to understand this is in Simondon's example, used throughout his work, of the difference between an air-cooled and a water-cooled engine. In the water-cooled engine, the water serves a single function in the "closed" field that is the engine. If we thought of things in this way, that is, if we thought of the engine as a closed system unto itself, we would forever be adding sections to it since to increase its operability we would have to consider elements that it does not contain, and then add them to it. But things do not really work this way. A different way to think about this is when we consider the air-cooled engine. In this engine, the air serves a variety of functions all without adding anything to the operation of the machine; it is an example of the technical artefact interacting with another "milieu" (as Simondon would put it). The information that air can change the engine not by addition but by introducing new informational properties that redefine the interoperability of the engine's components serves to introduce one moment of technological change. Simondon viewed all phenomena in the world in this way, including natural biological processes. Thus, it can be said that he sought a naturalized account of information (a notion that is still controversial today – see Chapter 3).

Simondon's sensitivity to contingency, lack of probability, and openness to the informational multimodality inherent in communicative processes are traits that he felt were equally important to the philosophy of information and to an understanding of its more phenomenological underpinnings. Indeed, he would take it one step further by introducing these features – which were until then associated with semantic information only – to information in the "hard" sense, that is to say, information as an *entity*. To put it in terms of a helpful distinction made by Floridi, information can exist in three ways: information "as" reality, information "for" reality, and information "about" reality. Where the cyberneticists thought the interoperability and indeterminacy of information "about" and "for" reality, Simondon thought these concepts in terms of information "as" reality. The key to Simondon's importance is his outlining of the metaphysics of information in a "hard" sense that remained "open." He provided a number of useful concepts and terms with which to talk about change and *information in action*. His concept of "disparation," for example, is used to describe the way two milieus (planes) of information interact with each other without ever really coming into contact (he uses the examples of left and right retinal imaging).

One of Simondon's contemporaries who deserves mention is another philosopher named Raymond Ruyer. Ruyer wrote on a great many topics far beyond the field of information, but it is worth noting that he penned one of the first philosophical treatments of this topic in his book (Ruyer, 1954). The text deals explicitly with cybernetics and information, but it also theorizes communication and reason by probing the nature of different types of machines, a field previously known as "mechanology" (this term is becoming popular once again). Ruyer, like Simondon, criticized the notion of automation, proclaiming instead that automation, rather than being a higher degree of perfection, actually reflects a certain disadvantage in the

form of information being cut off from the surrounding environment (again, what Simondon would have called a "milieu"). In short, the book represents a very early philosophical analysis of cybernetics and the philosophy of informatics, especially that of Shannon and Wiener.

The last of the early continental philosophers of information is Andre Leroi-Gourhan (1911-1986), who trained as a professional anthropologist. Another thinker of technology and information, Leroi-Gourhan wrote a book about technology titled *Gesture and Speech* (Leroi-Gourhan, 1964-65). In it, Leroi-Gourhan philosophized that the informational properties inherent in tools and different forms of technology actually serve as a type of extension of our minds. The book is filled with examples from his anthropological excursions where he unpacks the nature of tools, including bodily ones, and how they aided different societies mentally throughout history. He offers a variety of concepts to help him think through his findings, particularly in the section titled "The Expanding Memory". One of his most notable concepts is called "exteriorization", and is meant to describe when a piece of technology acts "in place" of a function that previously was only carried out by the mind. Leroi-Gourhan uses the intermeshing of tools and motive gestures, symbols, and our erect posture as examples, but today we could easily find many more, including but not limited to external hard drives, apps, and smart phones. Leroi-Gourhan was a philosopher-anthropologist who explained technology as exterior memory, and so even today some of his observations are completely relevant. Discussions about so-called "Big Data," memory implants, cloud computing, and wikis, are all completely in line with Leroi-Gourhan's early theories because he saw technology as a way to expand our knowledge and to act in place of biological memory. This can be also be seen as an early precursor to something like the "extended mind" thesis (Clark & Chalmers, 1998).

As we continue to explore our continental heritage in the philosophy of information, we will now turn to some of the later philosophers of information after Ruyer, Simondon, and Leroi-Gourhan. One of the more well-known philosophers in this tradition is Gilles Deleuze. Deleuze's contemporary Michel Foucault once said that one day the twentieth century might be referred to as "Deleuzean". Whether history will be kind to Foucault's proclamation or not is beside the point, but for the philosophy of information, Deleuze is certainly one of our key players.

Deleuze took many pointers from Simondon and, following his lead, expressed a politics of information along ethical lines. Though he is known much more for his work in ontology, a short text of his titled "Postscript on the Societies of Control" (Deleuze, 1990) was one of the first texts that articulated a (critical) political economy of information. In this short yet complex work, Deleuze ponders the relation of information systems to the function of society at large, and what control of these systems might mean (in the cold, hard, political sense) for notions such as democracy and politics. He coined the termed "dividuals" in this text (rather than "individuals") in order to express how human communication is divided among information systems. Alternatively, in some of his more metaphysical works, particularly in his magnum opus, Deleuze expanded on Simondon's work by investigating concepts such as the virtual/actual distinction (Deleuze, 1968).

Difference and Repetition is a dizzying book, due to the complex material and also Deleuze's opaque writing style. Nevertheless, the book's main thesis (that difference is *internal* to metaphysics; that it is not something between entities or representative of some type of unknown "gaps" in the world) popularized the notion, coming from Simondon, which holds that information really is the "stuff" of our world. What we perceive as "difference", then, is actually simply the interlocking of informational structures. Deleuze philosophized the virtual and multiplicity, control and communicability, all within this context of informational "immanence". Many since have followed Deleuze's lead and produced whole books on

"protocol" or "algorithmic culture", flat-ontology and information, adding to our understanding of the way that information can be used to understand the universe.

Continental philosophers after Simondon (such as Deleuze) added to our understanding of information. It should be noted, however, that while Simondon was working out of the tradition known as metaphysics, the other tradition from which many philosophers of information conduct their work is semiotics. While the continental philosophers of information concerned with metaphysics largely built on the ideas of Baruch Spinoza (1632-1677) and other European philosophers (Edmund Husserl, Martin Heidegger, etc.), those working in semiotics built on the ideas of the American philosopher Charles Sanders Peirce (1839-1914, see Chapter 6.6)) and the Swiss linguist Ferdinand de Saussure (1857-1913). For now, we will stick with Saussure, who seems to have made a greater impact with the European continental philosophers of information. Saussure's approach to semiotics was based on a theory that the notion of "meaning" consisted of three elements: a "signifier" (such as a rose) and a "signified" (love) which together produce a "sign" (rose + love). Some continental philosophers, Deleuze among them, thought that this distinction was too simple, and that the notion of difference, in a strictly philosophical sense, did not fit well with this semiotic distinction.

One of the first philosophers after Simondon who theorized information beyond semiotics from a variety of perspectives and in terms of how meaning changes throughout history was Michel Foucault. While Foucault wrote on a number of subjects – so much so that discussing his work could fill many volumes – some of the most significant concepts that he articulated in terms of the philosophy of information are "episteme" and "historical a priori". Foucault, more than any other of the continental philosophers of information, sought to think about information in terms of history and the archive. He asked such questions as "how does knowledge change throughout history?", "what is an archive?", and "how is meaning constructed socially?" He theorized the effects of information on things like society and the human body.

The ideas – episteme and historical a priori – systematically laid out in two major books (Foucault, 1966, 1969), are meant to denote the fact that anything we say and do, particularly when it comes to research, are in fact articulated within a field of "pregiven" information in the shape of discourses; that is, they are informed by meanings that have already been established. Foucault is thus most famous for analysing the social role of information for society and history, and even for the body (a practice he called "biopower"). For example, his research on the history of medical classifications helped show that certain fields related to illness and disease increased as classifications became available, thus enabling governing political powers to organize different groups of populations in certain ways ("madmen" in hospitals, lepers in "leprosariums", and so on).

Episteme and historical a priori, for their part, are meant to denote a certain "lower" level of meaning (Foucault often used the word "unconscious"), one that discursive fields depend on implicitly, without formally acknowledging their existence. Foucault often referred to the "positivity" of knowledge in this way, implying that certain knowledge is "spoken" and "made visible" through discourses, while deeper structures remained hidden beneath these discourses that are, essentially, taken for granted. This helps explain why Foucault was so interested in information in terms of archives, and why he wrote a whole book of original philosophy dedicated to the methodology of what he called "the archaeology of knowledge" that lay beyond positivism proper. Many have followed this lead. The philosopher Ian Hacking, for example, was inspired by Foucault when he wrote his well-known essay "Biopower and the Avalanche of Printed Numbers" (Hacking, 1982), a short text on the history of statistics.

Foucault understood information as a type of *representative* of organization – as an epistemology – and as what Foucault (1969) called "the system that governs the appearance of statements". In a very famous passage in the same book, Foucault says that individuals, people, networks, and texts are

> in a web of which they are not the masters, of which they cannot see the whole, and of whose breadth they have a very inadequate idea—all these various figures and individuals do not communicate solely by the logical succession of propositions that they advance […] they communicate by the form of positivity of their discourse, or more exactly, this form of positivity (and the conditions of operation of the enunciative function) defines a field in which formal identities, thematic continuities, translations of concepts, and polemical interchanges may be deployed. Thus positivity plays the role of what might be called a *historical a priori*.

When Foucault says that positivity plays the role of a "historical a priori" he is trying to get us to understand that we all play out the story of information, meaning, and communication according to determinate discursive fields. The important thing to take away from the passage mentioned above is that Foucault wants us to think of knowledge in informational (and, even more specifically, archival) terms where knowledge is less a progressive field of knowing according to successive inferences in the world than an architectonic system upon which we are able to express the current state of things via interventions and interpretations. In a way, he is suggesting that we cannot directly access something like "pure" knowledge or even history; that we do not communicate solely by the logical succession of propositions – this would exclude claims to universality made by certain positive discourses and in Foucault's time we can assume that he had certain linguistic theories in mind – and that we instead must play out the "positivity" of discourse even as we seek to circumvent, break away from, or create new styles of knowing. This is not some type of new age metaphysics or a type of analytic philosophy that theorizes the a priori in the vein of logical positivism. Put simply, Foucault, challenging the early Wittgenstein, is telling us that whereof one is silent, thereof one must speak. He is claiming that although we cannot but deal with the things themselves, we must do so in a new and interesting way that acknowledges their hidden underbelly.

This brings us to the next difficult philosopher of information, Jean Baudrillard, who is perhaps most closely associated with what some people refer to as "postmodernism". Baudrillard, the most controversial member of this group (just as Wiener was the most controversial member of the last), argued that the invention of information theory inaugurated the final division between the notion of "value" as tried to some physical thing and the notion of value for value's sake. He is the figure that is most associated with so-called postmodern philosophy due to his (at the time) esoteric theories concerning virtuality, simulation, and value. During his lifetime, many criticised Baudrillard's philosophy for being untenable, resembling science fiction, and seeming unrelated to contemporary concerns with culture and politics.

In retrospect, much of what Baudrillard had to say, from his neologisms concerning simulation to his philosophies on value, can now be understood as wholly in line with contemporary, albeit critical, debates on political economy. Perhaps the most well-known concept proposed by Baudrillard comes from (Baudrillard, 1976). Here, Baudrillard develops the concept of the "hypereality of floating values". The thesis here is that value is no longer tied down to anything "real" and that we have now entered an era of purely "symbolic exchange". Baudrillard wrote three other books on the subject of information, communication, and technology (Baudrillard, 1981, 1987, 2001), all of which theorized concepts such as "hyperreality", "simulation", "simulacra", and "symbolic exchange". Like Foucault, Baudrillard stems from the Saussurean branch of

philosophers who were more interested in deconstructing the semiotic apparatus of the sign in favour of a more abstract notion of meaning that was disassociated from any physical, tangible roots. Baudrillard also remained the preeminent postmodern philosopher of information for the way he problematized history in similar terms.

Finally, the last two continental philosophers of information on our list are the Stanford-based Jean-Pierre Dupuy and Michel Serres. They are the last of the living great French philosophers of information who are still publishing today. Dupuy (2000) is essentially an intellectual history of how Wiener and the rest of the cyberneticists laid the groundwork for cognitive science and AI, and also for modern debates on chaos and complexity theory, and how these relate to long-standing debates in the philosophy of mind. Serres, for his part, wrote a book (Serres, 1995) that is almost entirely devoted to the philosophical conception of "noise" as handed down to us by Shannon. Serres probed notions such as protocol, code, and atomism, as well as outlining different philosophical definitions of "noise", including "bruit noise" and "*belle noiseuse*" ("beautiful noise"). He analysed the different meanings of noise as a concept, and was a philosopher of early physics and atomism. Here we will end our early history of the cybernetic and continental versions of the philosophy of information.

1.8 Conclusion

While these uses of information are different from each other, they are all in some way related to the core idea of information separated from the peculiarities of the physical system sustaining that information at any particular place and time. And they all advance their fields by allowing concentration on what is happening to that information. This is what is fundamental to the information revolution: we can see things as similar that we never did before, while things that used to be separate – almost barricaded off from each other – can now be connected and interact.

In this chapter we have seen how vital the concept of information has become to the world, and understood where it has come from. These multiple profound new concepts originated in what were originally mathematical concepts. And note how Turing gave us only an idea – a new idea. He designed a new concept.

1.9 Exercises

1. Make a serious effort to imagine a world that had never thought of anything like Turing's universal machine. Now reconsider whether Turing's idea was mundane, or profound.

2. Imagine that we had Turing's idea, but still could not build a computer. How would the world have been different?

3. Try to think about different types of information. What would you call them? Can you provide examples?

4. Is information abstract or physical? Can it be both? Provide an example.

5. Are "meaning" and "information" separate? Why?

1.10 Further reading

Floridi (2011c, Chapters 2 and 3), Gleick (2011, Chapters 7 and 8), Wiener (1948, Chapter 8), Dupuy (2000, Chapter 1).

Chapter

2

2. WHAT IS THE PHILOSOPHY OF INFORMATION TODAY?

Information first

2.1 Introduction

> *'Evans had the idea that there is a much cruder and more fundamental concept than that of knowledge on which philosophers have concentrated so much, namely the concept of information. Information is conveyed by perception, and retained by memory, though also transmitted by means of language. One needs to concentrate on that concept before one approaches that of knowledge in the proper sense.' (Dummett, 1993, p. 186)*

Alongside academic philosophy, and the philosophical work commonly done in many other disciplines, Philosophy of Information (PI) is concerned with concepts. To communicate and co-operate successfully in pursuit of any goals, communities of people need shared conceptual schemes. When the world changes, because we find out about or create new things, those conceptual schemes need to change. For example, we needed to invent a new concept "spin" to develop quantum mechanics, and we need to understand how a Facebook "friend" is different from the usual kind of friend. For many working in PI, the purpose of philosophy is to help build the required conceptual scheme or schemes, to engage in conceptual design. And philosophy is most useful when it aims to build a conceptual scheme in response to the problems of a particular time and place.

In this chapter, we address the question of what PI is now by looking at what philosophical questions to ask today, and how to answer them. We will outline an approach to philosophical questions that influences a significant group of current philosophers of information – although it is not shared universally by current philosophers of information.

It seems to be human to ponder difficult questions. What is the soul? Is there a God? What is a lemon? Is there an *essence of lemony-ness*? There's nothing wrong with wondering about these kinds of questions. But if you want to do serious philosophical work, many current philosophers of information recommend that you think very carefully about your question, because some questions might be a waste of your time.

In this chapter, we examine how to discriminate between questions that are fruitful and questions that are not. It can require serious philosophical work to choose and refine a good question. One key idea to

grasp will be that we interact with the world at a particular Level of Abstraction (LoA), and failing to respect that can lead to a conceptual mess.

We will also begin to see why PI focuses so much on *information,* presenting the way Luciano Floridi, as a prominent current philosopher of information, sees PI as demanded by the "information revolution". The key idea is that, as our investigations of the world change, our understanding of both the world and ourselves changes too, and philosophy helps us come to terms with these changes, and to design new conceptual schemes to deal with them. We will explain how a vitally important recent change that demands new conceptual schemes has been the creation of the "infosphere".

2.2 The information revolution alters our self-understanding

It is not difficult to see that the progress of science has given human beings a radically different understanding of the world over the centuries. Could we ever have expected that time and space would be relative to a frame of reference? Who would have predicted the bizarre claims of quantum mechanics? Science changes our understanding of the world. But science does something else too – it changes our understanding of *ourselves* as human beings.

Science changing our self-understanding may be rare, but it can be a shattering experience. Consider three examples of scientific revolutions (as presented by Floridi, drawing on ideas by Freud) that were so profound that they altered forever how human beings see themselves and their place in the universe.

We might choose Copernicus as the standard-bearer for the first revolution, as he speculated that the earth was not the stationary centre of the universe, around which the planets, sun and fixed sphere of the stars revolved. Instead, the earth is just one planet of several in our solar system, itself orbiting the sun at an extraordinary speed. From a modern perspective it is difficult to understand how shattering it was to humanity's self-understanding when the empirical data eventually became so strong that the Copernican view became generally accepted as true. At least a large section of humanity had believed themselves to be so privileged that they existed at the very centre of a universe created entirely for them, so focused on them that it literally revolved continuously about them.

A natural standard-bearer for the second revolution is Darwin. While people had to accept that they lived on one speeding rock among many others, they considered themselves profoundly different from other animals on the planet. They were so clearly superior to other animals in their abilities to shape the world around themselves that they thought they must have had a quite different origin from even the most intelligent animals. Darwin destroyed all that. The theory of evolution showed how all life on earth could have descended from a few common ancestors, by gradual modification over many millions of years.

Recovering from this revolution with ill grace, humanity consoled itself with the view that people had changed so much since their common origins with animals that *now* they were different from other animals. They had rationality, and higher thought, which gave them independence from their baser emotions and instincts, quite unlike any other animal. Then humanity stumbled into the third revolution, spearheaded by Freud. He, and many thinkers since, convinced us that human beings are not entirely rational, nor are our minds perfectly transparent to ourselves. We do still have animal urges, and instincts, and sometimes our minds are opaque to us, so that we cannot always be sure why we are acting the way we are.

Together these three revolutions radically transformed our understanding of ourselves, to something much more like the current conception of humanity, and it is not possible to go back to how things were before. But science has not stopped. Our understanding of the world is still changing, and even the world itself is changing – because we are changing it.

Pioneering work on information by such giants as Turing and Shannon (see Chapter 1) led to the creation of information and communication technologies. The information revolution began with extraordinarily large, clunky computers in the 1950s, but it has recently exploded into a vast array of hardware that most people use every day (computers, laptops, tablet computers, MP3 players, e-readers, smart phones), and corresponding software, web services and apps that are also part of daily life (email, Skype, Facebook, Google). Without them, you could not be reading this e-book now.

The creation of the internet has had an extraordinary impact on human life, with many daily tasks (booking cinema tickets and flights, for example) and working life (email) transformed by it. The internet can be seen as the creation of an entirely new aspect of the world, accessed by newly available hardware and facilitated by the new software and services. This new aspect of the world might be called the information sphere, or *infosphere* (Floridi, 2011c, p. 14ff.).

Accompanying these dramatic changes in daily experiences has been the creation of vast amounts of data. Managing that data is now a large part of most jobs in the richer part of the world, and a large percentage of the GDP of advanced countries such as Canada, France, Germany, Italy, Japan, the UK and the USA is now made up of informational goods like music, novels, computer software, and other patented inventions such as drug formulae (as opposed to material goods that are made or grown in manufacturing or agriculture). This is why people talk of the new "knowledge economy".

So far all we have said is that the world and our understanding of it have changed. But revolutions of the kind we are interested in are revolutions so profound that they also change our self-understanding. Many philosophers of information believe we are currently in the throes of such a revolution: the information revolution. PI holds that we are coming to see the world and our place in it in a profoundly different way. We are coming to see that it is not just the internet that is the infosphere. The whole world is the infosphere, and we are informational organisms, or *infORGs* within it (Floridi, 2011c, p. 110ff.).

This is because our understanding of the world is changing from a world filled with very different and unique members of different kinds of things – the physical, like rocks and particles; life, like trees and rabbits; artefacts, like tables and cars. First, the boundaries between *kinds* of things are blurring, as we see their basic, most general nature as informational. Things of different kinds can interact, and even different kinds of interactions can be seen as fundamentally similar. The smashing of a rock falling off a mountain and the editing of software to change the action of a computer program when the "save" icon is clicked certainly have their differences. However, they can both be understood as informational interactions, changing data structures, as we have seen in the examination of the new language of information in Chapter 1, and as we will see in the discussion of naturalized information in Chapter 3. Second, the idea of a unique and irreplaceable individual object is also fading. An object like Michaelangelo's David is unique. Other objects very similar to it are not it, and do not share many of its properties – most notably its value, both aesthetic and monetary. But more and more our lives are concerned with non-unique objects. We no longer have to listen to a one-off, unique performance of a

piece of music, or buy a record that is initially identical to many others, but quickly scratches and acquires its own individual characteristics. We buy and download an MP3 file of a song, which remains interchangeable with all the other MP3 files of that song. This is the shift from material – unique and irreplaceable – to informational. In these ways, we are coming to see the whole world as the infosphere. Floridi calls this process *re-ontologizing* (Floridi, 2011c, p. 113ff.).

This new understanding of the world goes along with a new understanding of us and our place in that world. As in previous revolutions, the distinction between ourselves and others is being broken down. We are becoming freer of the constraints of our physical location and biological bodies. For example, our agency is no longer limited to our physical location, with most people having access to multiple means of acting at a distance, such as email, phone, Skype, Facebook, and so on. We cope increasingly seamlessly with multiple means of interaction, with interfaces like tablet computers and smart phones allowing us to move fluidly between distant action, and face-to-face action. Because of this, we can no longer see the internet and these multiple means of acting as alien and other, but as a seamless part of our lives, as just other methods for doing the things we normally do. There are fewer differences between processors and processed, online and offline. For example, a person driving using GPS is not clearly either online or offline. In these ways, categories developed in what we easily recognize as a new and special infosphere – the internet – are expanding naturally into the rest of our experience. Floridi calls this new status "onlife" (Floridi, 2011d, p. 550). It is because of this that we are coming to see the whole world as the infosphere and ourselves as inforgs.

This process is still happening. The current generation has not completely changed and can still see and experience the change. We are not yet fully-realized inforgs. But the babies currently learning to use iPads before they have the manual dexterity needed to turn the pages of a book will grow to become inforgs quite unaware of the change – they are the real children of the information revolution.

Just as for the previous revolutions, it isn't possible to go back. While we could lose the technology and have to return to a pre-Information and Communication Technology world, we will not soon lose the altered understanding of ourselves that the information revolution has created.

2.3 The philosophy of information as a field

Since the information revolution, there is a lot more information around than there used to be. And handling information takes up an increasing amount of time of increasing numbers of people, in both working and in leisure time. The scientific and mathematical study of information is now very important to the progress of information and communication technologies that affect all our daily lives.

As well as altering our self-understanding, this growth in information itself and in the science of information has done two things. First, it has opened up many more interesting problems, concerning what information is and how we are to understand it. Second, there has also been the creation of novel tools and methodologies ripe both for further conceptual investigation, and for plundering – to carry off novel concepts to help solve other problems.

This is what Floridi means when he writes:

> The philosophy of information (PI) is the philosophical field concerned with (a) the critical investigation of the conceptual nature and basic principles of information, including its dynamics, utilization and sciences, and (b) the elaboration and application of information-theoretic and computational methodologies to philosophical problems.
> (Floridi, 2010c, p. 46.)

We live in the "information age"; we are members of an "information society"; we read "information" in the papers; we can gather "information" on, say, the salt gradients of the currents in the Pacific Ocean; and we can talk about the amount of "information" that can be delivered over a wireless connection or stored on a memory stick. Because the term "information" is tossed about casually in common language with no clear meaning, it is useful to begin by clarifying its definition. So a core question for philosophy of information is "what is information"? Given that "information" is also used differently across different fields of study (biology, communications, computer science, economics, mathematics, etc.), it is a hallmark of the philosophy of information to undertake this task, if the term "information" is to be informative at all. So, first, on Floridi's understanding, PI is the research area that examines the concept and phenomenon of information in its many meanings and occurrences, and tries to clarify its many uses.

In asking the question "what is information?" this first goal of PI is a classic philosophical goal. The "what is *x*?" kind of question has been a core philosophical format since Socrates, who asked questions like "what is virtue?" and, "what is knowledge?" The aim of asking such a question is a unified theory of *x* – in this case, a unified theory of information. Although it is a classic kind of question, the "what is information?" question is timely because information is so vital to current work across the sciences and in mathematics and computing, so a unified theory of information would be of interest to many. Notice that it is worth *trying* to find a unified theory of information, even if such a theory is never found, as it is only by searching that we will realise there is no such theory. Philosophy has a role to play in attempting to construct such a theory.

PI is, however, much more than this. Given the information revolution and our radically changing understanding of the world as the infosphere and ourselves as inforgs, "information" is a foundational concept. So, some have attempted to overhaul philosophy by putting information first in philosophical enquiry, making the philosophy of information a new *prima philosophia* – a first philosophy. Traces of a developing field have been clear for some time, as described in Chapter 1. However, the most extensive overhaul of philosophy along informational lines comes from Luciano Floridi, starting with a series of articles beginning in 1995 and culminating (so far) in his 2011 *The Philosophy of Information*, published by Oxford University Press. In this book, he addresses several outstanding philosophical problems by borrowing concepts from the computational sciences and putting them to new use. Indeed, Floridi argues that there is a crisis at the heart of current philosophy that can only be addressed by looking at information as more foundational than the traditional categories of knowledge and existence. Other philosophers of information employ the notion of information simultaneously in biological, mathematical and philosophical terms, to address questions regarding the emergence of mind from a physical substrate up to and including consciousness. We will see how modern PI addresses many philosophical issues in the later chapters.

The second goal of PI is innovation: to develop an information-theoretic philosophical method in order to examine how ideas coming from the explosion of work on information might usefully be applied to philosophy, to develop new philosophical questions, and perhaps answer old philosophical questions. There are lots of ideas in the study of information, and it is worth thinking about what might be relevant

to philosophy. The results of this work to date will be presented throughout this book. We shall see in sections 2.5-2.7 that PI sees the method of Levels of Abstraction as one of the most important ways of importing information-theoretic ideas to philosophy.

Finally, it is worth noting that Floridi distinguishes between what he calls a "minimalist" and "maximalist" interpretation of PI. Floridi claims that everything is information. The whole universe and everything in it is made of information – the universe is the infosphere. Broadly, this is because information is the broadest Level of Abstraction (which we shall meet shortly) at which to describe everything. This may seem a bit much to swallow – especially if you are not familiar enough with Shannon and the multiple uses of information in science to have any idea why anyone would make such a claim! The maximalist approach to philosophy of information takes this idea seriously. PI is the philosophy of information design. As, according to this view, the world is made of information, this is a foundational philosophy. PI is the philosophy of everything.

However, accepting this claim is not necessary to find something interesting and important in PI. Information is clearly a very important concept, and it is one that has been neglected within much of academic philosophy. There are conceptual problems arising in ICTs. Even if the whole world is not the infosphere, the internet is. And the internet and related technologies and how we use them really are revolutionizing our lives. The conceptual investigation of information, and application of information-theoretic methods, is still very interesting. The choice of maximalist or minimalist approach to PI is up to you. There is a great deal to discover either way.

2.4 Open and closed questions

'It is precisely this that marks out a problem as being of the true scientific spirit: all knowledge is in response to a question. If there were no question, there would be no scientific knowledge.' (Bachelard, 2002)

Floridi (2010c) identifies and discusses open problems in detail. What we examine here is the *idea* of an open question. This idea is important to PI, and can be useful to you in any philosophical (indeed any academic) studies.

In brief, the idea of an open question is a question that is open to informed, intelligent and reasonable disagreement. You can see that quite a lot of questions are open: how should Alice wear her hair? Which film should Alice and Bob go to watch? Which party should Carol vote for? Interesting open questions are about subjects of current concern. They should be precise enough that we can expect substantive progress within a reasonable time period. A reasonable time period depends on how important the question is, and how long it is likely to remain important. But as a rough guide PI is looking for progress in less than a century – and hopefully less than a decade! An open question is also related to the idea of Levels of Abstraction (LoAs). An open question is not asked independently of specifying its LoA, which we will examine in sections 2.5-2.7. An open question is framed for a purpose, in the context of consideration of what kinds of things are relevant to answering it.

This is in contrast to a closed question. A closed question is one that can in principle see disagreement resolved once and for all, by some facts, or some calculations. A bad question – whether open or closed – is unanswerable not because we may disagree on the answer, but because there is no satisfactory way of

both accepting how the question is framed – including the assumptions that are implicit in the question – and answering it on its own terms.

So the first step in identifying good open questions is to ask whether they are of current concern. Do they matter? For example, "What is the soul?" might be a question of current concern, as is the question "Can machines have souls?", at least in the sense that many people currently care about these questions. An answer to these questions might be important. In contrast "Is there an essence of lemony-ness?" seems less important, and actually concerns very few people. We abandon that question as failing the first step.

The second step is to ask whether the question is answerable. "What is the soul?" and "Can machines have souls?" are difficult. How are we supposed to go about answering them? What are the considerations that we bring to bear on deciding how to answer them? These questions, in these forms, are vague. One at least is a question that philosophy has been addressing for millennia – without a great deal of progress. This is a reason to proceed cautiously, to ask yourself whether you can contribute something new. These questions fail step two.

You may reject the initial question entirely if it fails step two. As an example of a bad question to help guide your understanding, consider "what is the real nature of buggles?" This is clearly unanswerable for you, as you have no idea what a buggle is, or why anyone should pose such a question, and so no idea how to begin to answer it. More seriously, the debate over scepticism, which you may know, has been running, without satisfactory answer, for millennia. Floridi argues that it is a bad question, forever unanswerable, because if you accept the requirements for an answer, you will see that they cannot be met. Floridi also argues that the Gettier problem cannot be answered, unless the way the question is framed is changed. This is examined further in section 8.3. It is possible that rejecting a question as bad is the right path to take if it fails step two.

However, there is a third step in identifying good open questions. It is to wonder whether the original vague question can be reframed in more specific terms – turning an unanswerable question into an answerable one. Questions can be made more specific – refined – in multiple ways. The best way to choose a more specific question is to think, "*why* does the question matter?" and, "why does it matter *now*, and is this different from why the question mattered – if it did – a decade ago, a century ago, and a millennium ago?"

After asking why it matters, we might change the question "What is the soul?" into "Can we conceive of the human psyche mechanistically, so that we can understand how the mechanistic accounts of various cognitive tasks of *parts* of the psyche that psychology, cognitive science and neuroscience attempt to provide relate to the nature of the *whole* psyche?" There may be many ways to create more specific questions; this is merely one example. But it is timely, and open. It is framed in terms of a particular purpose, and indicates relevant considerations for answering it, and literature to draw on. It is seeking to reconcile the mechanistic approaches of (some) sciences to (some) explanations of (some) human abilities to the way ordinary human beings conceive of human mental life. There is already a great deal of philosophical work on understanding the elements of this question, and there is source literature that makes it possible to understand them, and to understand what an answer to the question means. Like any philosophical question, better formulation of the question continues to be a concern in attempting to answer it. But this question has already made an important step away from the vagueness of "What is the soul?"

Take the second question "Can machines have souls?" In trying to make this question more precise, we might come to the question "Can machines think?" One of Turing's profound contributions to thought is his revolutionary recasting of an unanswerable question in answerable terms. Lost in the complexities of Turing machines (see Chapter 1), it is easy to forget that Turing refused to try to provide an answer to the question "can a machine think?" because he considered it a problem too meaningless to deserve discussion. He objected that the question involved vague concepts such as "machine" and "thinking". So he suggested replacing it with the Imitation Game, which is more manageable and less demanding because it fixes a rule-based scenario that is implementable and controllable. The basic idea of the Imitation Game is to provide a test for, rather than a definition of, consciousness – or some aspect of consciousness. Turing's idea is that we have reason to believe that machines can think, if a person asking questions – say over the internet – cannot tell which of two respondents is the person, and which is the machine (Turing, 1950).

Turing's idea of this kind of test provides a new question that contains specification of what is to count as an answer. Turing chose to compare "machine intelligence" to human intelligence. In testing whether computers can imitate us in answering questions, we are testing only whether machines can *answer questions like us*. The comparison could have been to something else, from animal intelligence to human creativity, as many other versions of the Turing imitation game have shown. What is important is that Turing asked a new question, which may be summed up as "may one conclude that a machine is thinking, by finding that one cannot discriminate between human beings and machines in the imitation game?" After half a century, philosophy is still learning the crucial lesson that *asking the right question is a vital part of the philosophical battle*. This is one of the greatest and lasting contributions of Turing's famous test, far more important than the inaccurate predictions about when machines would pass it, or what conclusions one should draw if they did. Other thinkers through the ages have pressed us to rethink questions – often thinkers who were subversive and destabilising in their own time, with their true power being realized only later. Within philosophy, one example is Wittgenstein.

The difference between open and closed, good and bad questions is very tricky. The only way to get it really clear is to look at philosophical questions for yourself, and start trying to assess whether they are open or closed, good or bad. We will also find out more about how to frame good philosophical questions in sections 2.5-2.7 below. So the best idea is to bear the distinction between open and closed and good and bad questions in mind while you read the other chapters of the book. For example, we will see how PI's approach to ethics (Chapter 4), knowledge (Chapter 8), and personal identity (Chapter 15) differs from traditional approaches. PI aims to identify good open questions in traditional philosophical debates, and to avoid bad or closed questions. Think about which kinds of questions you prefer. It might help you to re-read this chapter after having really studied some of the others.

2.5 The idea of a Level of Abstraction (LoA)

Take a blank sheet of paper and a pen. Take five minutes, and observe. Really, try it. No, we mean it.

If you really tried it, you will have come to understand something very important. Think about what you wanted to ask. To help you complete your task, you almost certainly wanted to ask two questions: Observe *what*? *Why*? You wanted to know *what* things you were supposed to observe, and what things to ignore. Am I to observe the people in the room? What about them: is it their behaviour, their height, their gender ratio, or what? Or am I supposed to note the décor, the temperature, or the function of the room? Knowing what to observe might have allowed you to guess *why* you were observing. Am I investigating the ratio of male to female students on philosophy courses? Am I worried about funding cuts to universities leading to declining facilities? Am I worried that the heating is broken and students can't learn effectively if they are too cold? If you know *why* you are observing, this will help you guess what to observe.

For many philosophers of information, we need answers to these questions whenever we interact with the world in any way – when we think about it, talk about it, look at it, much less *do* anything to it! We don't usually make explicit the answers to what we are observing and why, but nevertheless assume them when we think, talk, look, or act. Details about what to observe are Levels of Abstraction (LoAs) and making them explicit avoids mistakes.

Suppose Alice and Bob are leaving a party, and Alice tells Bob her address is "56B Whitehaven Mansions, Charterhouse Square". Bob makes assumptions about that address. Since they are in Oxford, he assumes it is an Oxford address. But later when he tries to check it on an online Oxford map, no address is found. Bob makes the false assumption that the right LoA at which to consider that address is as an Oxford address. The same happens if Bob switches his LoA to "England", and then to "UK". Eventually, Bob Googles the address. Now the LoA has completely changed. It is wider in scope, and allows many more types, which is just to note that now Google will produce lots of different kinds of information online related to that address. The first entry makes Bob feel like a fool: "56B Whitehaven Mansions, Charterhouse Square" is indeed a place in Smithfield, London W1. But it is the address of a fictional retired Belgian police officer, Monsieur Hercule Poirot. Alice unkindly used the LoA of a novel to mislead him.

The idea of LoAs is crucial to handling any information process, and so to how we think about and perform our interactions with the world, and therefore in how we develop our philosophy of information. *LoAs are important whether they are made explicit or not*, as we see in Alice's reply to Bob. Because of this, the only way to avoid mistakes in important cases is by making LoAs explicit.

The Method of Abstraction comes from modelling in science, where the "variables" in the model correspond to the things chosen to be observed in reality. The term "variable" is commonly used throughout science to stand for an unknown or changeable value of something measured. So people's height is a variable that can take many values we can measure, such as 152cm, 163cm, 1.6m, 5 feet 5" and so on. The variables are measured; everything else is ignored. The choice of variables – and the implicit choice of what to ignore – depends on the purpose of the observations, and in the end, the model you are making. The terminology of LoAs has been influenced by an area of computer science called Formal Methods, in which discrete mathematics is used to

specify and analyse the behaviour of information systems. Despite that origin, the idea is not at all technical and, for the purposes of this chapter, no mathematics is required.

2.6 The definition of a level of abstraction

Suppose we join Alice, Bob, and Carol earlier on at the party. They are in the middle of a conversation. We do not know the subject of their conversation, but we are able to hear this much:

- Alice observes that its (whatever "it" is) old engine consumed too much, that it has a stable market value but that its spare parts are expensive.
- Bob observes that its engine is not the original one, that its body has recently been re-painted but that all leather parts are very worn; and
- Carol observes that it has an anti-theft device installed, is kept garaged when not in use, and has had only a single owner.

Alice, Bob and Carol view whatever "it" is according to their own purposes, which guide their individual choices of what to pay attention to, their LoAs. We may guess that they are talking about a car, or perhaps a motorcycle, but it could be an airplane, since any of these three objects would fit the descriptions provided by Alice, Bob and Carol above. Whatever the object is, it is the source of information under discussion. We shall call it the *system*. But they talk about it differently. Perhaps Alice is an insurer, Bob tinkers with engines, and Carol is a collector and potential buyer. They all view the object at different LoAs. An LoA consists of a collection of observables, each with a well-defined possible set of values or outcomes.

This is what Floridi means when he gives a more formal account of levels of abstraction:

> A level of abstraction is a finite but non-empty set of observables possibly moderated by transition rules.
> An observable is a typed variable with a label, that represents the name assigned by the epistemic agent to a feature of the system under consideration.
> A typed variable is (i) a place-holder for an unknown or changeable referent; and (ii) a set, called its type, that consists of all the possible values that the variable may take.
> A transition rule is a predicate that provides the trajectory of change of the observables inside its type. (Floridi, 2010c, extracted from Chapter 3)

We can state the key idea the simple way: people using different LoAs concentrate on different features of the object, observe those features, and so describe the object very differently. Or we can say the same thing more scientifically: people using different LoAs concentrate on different features of the system, different observables, choosing different variables to measure those observables, and so model different aspects of the natural system.

Variables are used ubiquitously in science. But the example above can also show how using variables to track observables might be of interest even in informal conversation. Suppose Alice, Bob, and Carol are talking about a plane. Alice's LoA might consist of observables for running costs, market value and maintenance costs. Suppose Alice insures the object. If she wished to support the claims she makes about it, she could choose variables to measure features of, say, running costs, such as a variable to measure the engine's fuel consumption. Bob's LoA might consist of observables for engine condition, external body condition and internal condition. Suppose Bob is a mechanic, in which case his job will require him to

pay attention to some rather more precise variables, such as day to day oil and fuel levels. Carol's LoA might consist of observables for security, method of storage and owner history. Suppose Carol is the owner. She, too, might choose variables to measure some of her observables accurately. Each of Alice, Bob and Carol's LoAs makes possible a determinate analysis of the system, and we call what such an analysis yields a *model* of the system.

So it can be seen how different purposes lead people to pay attention to some features rather than others of the same system. The observable is simply something that the observer chooses to pay attention to, while variables are used to measure more accurately the state of that observable. Note that you had better pay attention to the right features for your purpose! If Bob, as the mechanic, worries mainly about the shiny paint, sooner or later the plane will crash. Note also that, since the system investigated may be entirely abstract or fictional – recall the example of Poirot's address – the term "observable" does not mean "empirically perceivable". Ultimately, an *observable* is just an *interpreted typed variable*; that is, a typed variable together with a statement of what feature of the system under consideration it represents e.g. a set of data could have millilitres as a type and oil level as an observable of the system.

This brings us to typed variables. Everyday conversation seldom requires typed variables, but from time to time this need can intrude even into informal conversation. Perhaps Alice and Bob got into a furious argument over the value of the plane, because Alice, being British, was using pounds, while Bob, being American, was using dollars. Had they specified the type of the variable, they could have avoided an entirely useless argument – and Alice may have given Bob the right address! In science, typed variables might not be explicitly stated if there is a standard unit of measurement always used in a particular field. But NASA lost the 1999 Mars Climate Orbiter because the US team used the metric system of units of measurement standard in the US, while the UK team used the imperial units of measurement standardly used in the UK. Leaving types of variables implicit can cause mistakes.

So far we have only worried about looking at a system at a particular time – measuring a set of variables. But things change over time. Often science has to pay attention to how sets of variables measuring observables of a system change over time, so that they can learn how the system changes over time. That is what transition rules are, telling you how one or more observables change over time. The most familiar example in our case is fuel levels in the plane. If the plane is unused, they remain stable. As the plane is flown, fuel levels can be expected to decline steadily. When the plane is prepared for flight and re-fuelled, fuel levels jump quickly from low back to high again.

The same object, such as the plane, can be viewed from many different LoAs, and they might be related to each other. Floridi calls a relation among LoAs a "Gradient of Abstraction" (Floridi, 2010c, see Chapter 3). The relation can vary a lot along a continuum, but there are two ends of that continuum. Two LoAs are *disjoint* when they are just unrelated, even though they are about the same object. We can state this precisely now: two LoAs are disjoint when they share no observables. So for example, a child's view of the plane might consist of only "It goes so fast and it's so *loud*!" The child's LoA has only two observables: fast and loud. This might share nothing with Bob's far more complex LoA, containing many observables necessary to a mechanic. At the other end of the continuum, two LoAs are strongly connected when they are *nested*. This is the case, for example, if we suppose that Carol, the owner's, LoA for the engine consists of only one observable: engine condition. Her variable to measure engine condition may be very simple, so suppose it takes only three values: good, ok, and bad, depending entirely on the opinion of Bob, who we now suppose is her mechanic. Then all of Bob's more detailed variables

for aspects of the engine's condition are all related to – indeed, they generate – Alice's cruder choices of good, ok, and bad.

In summary, we have seen that in spite of the technical language used to describe them, the core idea of LoAs remains a very simple one: whenever we think, talk, look, or act, we interact with the world for a particular purpose, paying attention to certain features of the world, certain observables, and completely ignoring other features. We may wish to specify variables to measure observables more carefully, and giving their type and transition rules for how they change over time can also be very helpful. In doing this we build a model that extracts some information from the world, and in the process of doing that, ignores a great deal of other information.

2.7 The implications of LoAs

> *'But do not ask absolute questions, for they just create an absolute mess.' (Floridi, 2010c, p. 149.)*

The simple fact that we always interact with the world via an LoA has implications for many debates in philosophy, and we will see this fact playing out in many later chapters of this book. Here, we focus on two main ideas. First, we will see that LoAs imply a certain kind of *pluralism*, but only a very modest kind, rather than a form of relativism. Second, we will make explicit what the discussion of LoAs should tell us about avoiding useless questions. The first involves taking care not to overestimate the implications of LoAs, while the second involves taking care not to underestimate their implications.

We hope it is quite clear by now that there can be many different ways to look at the same object. There were many different – and perfectly sensible – LoAs at which to consider the same plane described above. This is a kind of pluralism. But it is a mistake to think that any LoA would do. Consider Bob, the mechanic, choosing to use the child's LoA: "it's *fast*, and *loud*!" The plane would soon crash, as Bob ignored vital observables. Bob is forced to choose observables, and variables and their types, appropriate to his purpose. As an insurer, Alice should pay attention to some of the same observables that concern Bob, but she will soon lose money, and her job, if she does not also pay attention to things Bob might try to hide from her, like frequency of maintenance, and also things that are of no interest to Bob, such as anti-theft devices on the plane. LoAs may be relative to purpose, in the sense that it is your purpose or goal that guides your choice of observables, but they do not imply any arbitrariness in the subsequent choice. To say the same thing more scientifically, different interfaces – LoAs – may be correctly ranked depending on how well they satisfy modelling specifications such as informativeness, coherence, elegance, explanatory power, consistency with the data and so on, and the purpose orienting the choice of the LoA. Recall that even in cases where the LoA is usually left implicit, there can be good reason to assume a particular one. Bob was perfectly entitled to assume that the address Alice gave him was an Oxford, or at least a real UK, address. Since Alice knew that he would make that assumption, her game with the LoA was not a clever trick, but a cruel deception. NASA's Orbiter crashed because the types of the variables for a key component were left implicit. It would not have crashed if the units were made explicit, or if metric units of measurement were standard worldwide.

The modest nature of the pluralism above has implications for what we conclude about the world, when we model it using LoAs – which is to say, it has *ontological* implications. A theory commits itself ontologically by opting for a specific LoA, because by adopting an LoA a theory commits itself to the existence of some specific

types of observables characterising the system and constituting the LoA. Consider choosing a car to buy. At the stage at which you have decided to buy a Honda Jazz, but you haven't bought one yet, you are committed to the existence of that *kind* of car. Next, by adopting the models that the LoA allows, the theory commits itself to the corresponding tokens. You are committed next to the factory having managed to produce some actual physical cars of that kind (that make and model). Finally, a theory is ontologically committed in full by its model, which is therefore the bearer of the specific commitment. At the stage of buying a particular car, which is an instance of the type 'Honda Jazz', one commits oneself to the existence of that particular car, light blue, with a particular registration number, parked outside your house. You gradually commit yourself to there being various things in the world that support the model you are building.

So while there are lots of ways of looking at things, each in some way commits you to there being something that you are looking at or talking about or interacting with. This is why your LoA or LoAs cannot be arbitrarily chosen. If you do not choose well, the model you are building of the world will cease to help you. Try going to a car dealership and asking to buy a spaceship just like the *Millenium Falcon* – unless they have the new model *Starship Enterprise* from the reboot universe available, of course. Ultimately, LoAs have modest pluralist implications for our view of reality.

It remains only to spell out what LoAs tell you about philosophical questions, and here it is important not to underestimate the importance of LoAs. Recall from section 2.4 that PI recommends looking for open questions, questions that are timely and important and phrased in such a way to be answerable. PI recommends rejection of philosophical questions that don't matter, or are out of date, or unanswerable. We are now in a position to specify more carefully one kind of surprisingly common unanswerable question: an absolute question.

An absolute question is a question that demands an LoA-independent answer, such as: "What is the One True Essence of Lemony-ness?" or, "What is the Real Ultimate Value of the plane?" These are a problem because, if we always interact with the world at an LoA*, there is never an LoA-independent answer to any question!*

This is why Floridi says that to ask absolute questions creates an absolute mess. It is because only questions asked within a shared LoA can be effectively answered. Suppose there is an argument about the value of our plane. The participants need to specify whether they are talking about monetary value, aesthetic value, or functional value for travel, for example. Once they have decided that the argument concerns the plane's monetary value, they still need to type the variable by specifying what currency they mean. Care might even yet be required. For example, is the issue what price the plane should be to buy, or its insurance value? They can be different. Until you have settled these questions, there can be no meaningful discussion. Any discussion without these questions settled is probably a waste of everyone's time. This is what is meant by saying that only questions asked within a shared LoA count as open questions, because only these have a possible answer. According to PI, these are the kinds of questions philosophy ought to concern itself with.

We can return briefly to the examples used in section 2.4 to illustrate how to find open questions, or refine existing questions to make them open. The first example was: "Do people have souls". At stage one, we said that it matters to many people now, so it could count as of current concern. But as it stands, it fails stage two, because it is so vague we can't see how to answer it. Reconsider the questions we asked at stage three in order to make the question more specific. The questions we recommend you ask yourself are: "Why am I asking this?", "Why does it matter?" and "What considerations would help me decide one way or the other?" You can now recognize these as precisely the questions aimed at making explicit the

purpose of the LoA you are trying to specify, and home in on what aspects (of, in this case, people, and in the earlier example, the plane) to pay attention to, to decide the question.

The second example of section 2.4 was "Can machines think?" We said that this is of current concern, but noted that Turing rejected it as too vague to answer. It was certainly much more difficult in Turing's time, before recent developments in computing and AI and so on, to try to specify relevant observables and variables to measure them. We had no idea what kinds of things might be useful to observe! An important aspect of Turing's imitation game was to condense possible observables into a much simpler test – can we tell whether the thing answering our questions over the internet is a person or a machine? This is a dazzling trick, condensing a lot of possible observables into one: can you tell the difference, which can only take two values of yes or no. The genius of this is that you hope that what you ignore – in this case because you don't know about it – won't matter. So Turing showed that you can be quite inventive in changing a bad question to an open and good one when you have a problem specifying the LoA! In short, LoAs help you diagnose questions that aren't worth your time, and they also positively guide you in creating useful open questions.

We have finished the extended introduction to what philosophy of information is now. We are in a time of rapid technological change, which is straining our shared conceptual schemes. Philosophy – in the academy and elsewhere – has a great deal to contribute in designing concepts for our time. The information revolution means that much of the conceptual strain is related directly or indirectly to information. In this chapter we have seen why it is important to be careful about which philosophical questions to devote real effort to. LoA-independent questions might be fun to worry about, but they are not likely to reward hard work with fruitful progress. But understanding LoAs also helps in a positive way to get questions into good order so you can seriously answer them. They should help you take those philosophical irritations, those issues that really get to you, and start making serious progress on them.

2.8 Exercises

1. Think about your technological ability and compare it with the technological ability of an elderly relative. Do you have a grandmother who doesn't know how to text, for example? What do you think might be different in the way you understand *yourselves*?

2. Imagine what kind of technological abilities infants learning to use iPads now will have once they reach adulthood. Do you think they will understand themselves differently?

3. Pick a philosophical question that really bothers you, and work through steps 1-3 described above. Do you think you have refined the question in a useful way?

4. If you abandoned the first question after steps 1 or 2, try again with another!

5. Describe the same object at three different levels of abstraction.

6. If you are familiar with the concept of a computer interface, think about what is the difference, if any, between a level of abstraction and a computer interface?

7. Does the adoption of the method of levels of abstraction commit you to some form of relativism?

2.9 Further reading

Chapter 1 of Floridi (2011c) expands on what is philosophy of information; Floridi (2010c, Chapter 2) examines philosophical questions; and Floridi (2010c, Chapter 3) explains LoAs.

Chapter

3

3. NATURALISED INFORMATION

Objective aspects of information

3.1 Semantic vs. natural information

> *'The higher-level accomplishments associated with intelligent life can be seen as manifestations of progressively more efficient ways of handling and coding information. Meaning, and the constellation of mental attitudes that exhibit it, are manufactured products. The raw material is information.' (Dretske, 1981, p. vii.)*

When we say that someone becomes informed of an event, say by receiving an e-mail containing the news of a successful job application, we would most likely construe this process as the recipient's learning something from the message. The sender – the possible new employer – wants to convey the news to the applicant, and in order to achieve that, she makes use of a language that is understood both by her and by the applicant.

Typically, a language is quite an arbitrary construction. Both in our written and spoken language, a word and its corresponding object hardly resemble each other. Still, the information is successfully communicated, because the receiver knows the general meaning of the language's symbols, which enables him to access what is called the *content* of the message. The content of a message is not what you actually see when you read something, but the things the message *refers to*. Content is *extrinsic* both to the collection of signs of a message, and to their physical carrier, which can be dots on a printed piece of paper, or the collection of pixels on a computer screen. The sender encodes the content of a message before sending it, and the receiver decodes the information after receiving it, thereby getting access to the content.

When asked to give an example of information transfer, a lot of people would probably think of a situation similar to the one involving the notification about the outcome of the job interview via email. Such examples will likely involve persons who communicate via languages like English. They will involve information that is cast into a symbolic form in order to be communicated. The persons who make sense of these symbols are called "semantically enabled agents", since the semantic aspects of the message, or, what the message *refers to,* is of concern to them.

However, we also observe that the use of the word "information" is not confined to such cases involving semantically enabled senders and receivers of messages. For example, we speak of information being transferred when referring to the exchange of data between fax machines, or between clients and servers

that transmit data by means of internet protocols. In both cases, the machines involved do not understand the information in the way we do. Moreover, we use the word "information" even for systems that are not constructed by human engineers, like the transfer of genetic information between chromosomes of the cell nucleus and the ribosome that produces the polypeptides according to the information encoded in the messenger RNA. In a wider sense, the fact of the past existence of animals can be communicated to a sedimentary layer, which preserves this fact as a fossil, or, in other words, as a record, which in turn can be considered as a kind of information. Likewise, the life story of a tree is communicated to its rings, and thereby they hold information about the tree's age. In these examples there are no senders and receivers making sense of symbolic language, but we still comfortably describe the examples by referring to "information transfer".

The naturalization of information is the research programme that tries to explain how the transfer and utilization of information is possible, where the reference of information comes from, whether concepts like meaning and content, and ultimately mental concepts like belief and knowledge, can be based on information. Crucially, by developing such an account, we do not want to presuppose the existence of semantically enabled agents, which can make sense of information by virtue of "understanding the language" into which the information is cast. If there is natural information, and if systems devoid of understanding can transfer and store information, then it does not solely reside in the heads of human beings, but has a status of its own. We examined this idea initially in Chapter 1, introducing Alan Turing and Claude Shannon, but here we examine it in much more detail. The chapter's main focus will be on the transfer of information and aspects of its semanticisation, but not on further aspects of natural information processing. (For information processing, particularly the concept of a Turing machine, see Chapter 13 on computation.) Also, this chapter will consider the relevance of natural information to mental concepts only in so far as it is relevant to understand how it gives rise to semantic structures. For more detail about the mental aspects of information, you might want to look at Chapter 10 on cognition and Chapter 11 on mind.

One of the central questions concerning the transfer of natural information is the problem of reference: how can one physical structure refer to what has happened with another physical structure at a different time and at a different place, so that information is retained while travelling through time and space? If something like this is possible, one can argue that more than the current physical state of the universe exists. A subsequent question concerns what else is needed for the process of "becoming informed", next to the fact that the information has to travel from the source to the receiver. We will see that this next step is fundamentally different from being a kind of transfer. The chapter will look at three different accounts of naturalization of information. Fred Dretske, the pioneer of this project, will receive the most attention, and his account will be contrasted with that of Radu Bogdan. Finally, we will look at another famous "naturalization" project, Quine's naturalization of meaning. Understanding the fundamentals of this problem is beneficial to appreciate current debates on neuro-physiology ("When does the transduction of information stop and when does the interpretation start within the brain?"), artificial intelligence ("Can we build semantically enabled artificial agents?"), and metaphysics ("Is information the ultimate substance of the universe?").

3.2 Information channels and semanticisation

We will see that a concept that is crucially connected to naturalized information is so-called "environmental information", which is a particular coupling of two systems connected by a channel that can be considered as a conceptual extension of the Shannon communication channel, as introduced in Chapter 1, and described in (Shannon, 1948). Shannon-information is encoded and decoded: it is transmitted through a physical channel, the way a mobile phone signal is sent, and sender and receiver are usually situated in different places. But these constraints are not necessary for the idea of information transfer – what is necessary is that information is communicated through time to a system that subsequently holds the information, as with the case of the animal and its fossilisation. And as the tree ring example shows, the systems communicating do not even have to be distinct.

The first step in the naturalization of information therefore is to identify the physical process that produces the event at the receiving end of the channel construed in this more general sense. The second step consists in the subsequent interpretation of the information that is held at the receiving end of the channel. This is what is called the *semanticisation* of the information: the information comes to have meaning. Think about the example of becoming informed about the outcome of the job interview: you open the email on your computer screen and the pixels assume a certain pattern of combinations of black and white. So far, all the information generated at the source is preserved. But there will be a point when this ceases to be the case. Maybe you will browse the content of the e-mail quickly in order to get to the important bit about the employment decision, and this will trigger a further reaction of yours. This process is an extraction of one part of the total information contained in the message; the rest will have been dropped and ceases to exist if the e-mail is deleted after reading it.

3.3 Environmental information

Before turning to some details of semanticisation, let's look at the relevance of environmental information to the naturalization of information:

> Approaches to semantic information also seek to connect it to other relevant concepts of information and more complex forms of epistemic, mental and doxastic phenomena, in order to understand what it means for something, e.g. a message, to be informative. For instance, Dretske (1981) and Barwise and Seligman (1997) attempt to ground factual semantic information in environmental information. The approach is also known as the naturalization of information.
> (Floridi, 2011c, p. 54.)

In this formulation, "environmental information" is to be understood as follows: 'Environmental information =def. two systems a and b coupled in such a way that a's being (of type, or in state) F is correlated to b being (of type, or in state) G, thus carrying for the observer of a the information that b is G.' (Floridi, 2011c, p. 32.)

So for example, the tree trunk having a certain number of rings (a's being F) is correlated with the age of the tree, so that an observer who counts the rings also knows the tree's age (b's being G). Alternatively, a ribosome produces the polypeptides from those amino acids that correspond to the base pairs that the messenger RNA carries.

But how is this coupling connected to the information transfer through a Shannon channel? Note first that successful information transfer does not depend on a preceding agreement on a code. The information resides in the environmental states of the carrier, not in the semantics that are assigned to these states. That is why the coupling of the systems, constrained by the formula for environmental information, fulfils its role as a medium for information transfer. However, the formula conspicuously contains a reference to an "observer" – probably a semantically enabled one. Environmental information covers the first part of the naturalization, the natural transfer of information. The tree's age is communicated to its own physical structure. But an observer must still interpret the structure to extract the information. In order to succeed in that, he must know about the correlation between age and number of rings. What the extraction of this piece of information amounts to will be considered a bit later. For now, the difference between environmental information and semantic information as ordinarily understood is of relevance.

3.4 Equivocation

We can start by looking at an example of environmental information in human behaviour, which shows that even semantically enabled agents, while communicating, do not necessarily produce semantic information:

> On the other hand, an event or state of affairs that has no meaning in any conventional sense may carry substantial amounts of information. An experienced poker player can read the signs; he can tell, or be reasonably certain, when a bluff is in progress. His opponent's nervous mannerisms, excessively high bet, and forced appearance of confidence reveal, as surely as would a peek at his last card, that he did not fill his inside straight. In such a situation information is communicated, but the vehicle that carries the information (opponent's behaviour and actions) has no meaning in the relevant conventional or semantic sense.
> (Dretske, 1981)

Having noted that pre-assigned meanings are inessential to information transfer, it must be noticed next that if the observation that a is F, such as the player's excessively high bet, carries the information that b is G, that a bluff is in progress, it is implied that the coupling of the two states is unequivocal. This is a crucial feature of Dretske's interpretation of a Shannon channel. Equivocation occurs if several different events at the source are conflated into one observed event with the receiver, thereby rendering it impossible to infer exactly which event at the source produced the event with the receiver. For example, if the player's excessively high bet could be caused by a lapse in concentration, then perhaps the player is not bluffing. This idea constitutes the difference between a causal transfer and an information transfer. Every information transfer is a transfer of a causal influence, in the sense that the event observed at the receiver is brought about by the event at the source. (This is a simplification of Dretske's view. We do not need the complexities for this introduction.) Whether the high bet is caused by an attempt at bluffing, or by inattention, it is caused by something, so these are causal transfers, even if the equivocation – the inability to tell *which* is the cause – means that it is not an information transfer. So, not every transfer of causal influence is an information transfer. On the other hand, we also have to contrast systems that are unequivocally correlated with informationally coupled systems. This contrast will be dealt with again during the discussion of Dretske's intentional states. At any rate, it should be clear by now what it means for environmental information to pass through a channel that is equivocation-free: it means that information is not irreversibly lost. In the poker case, it is the combination of nervous mannerisms, forced confidence, and the high bet that makes the experienced player sure that his opponent is bluffing.

In what we have considered thus far, we are still looking at a case where a signal has been transferred through time to a system that holds the information in the signal in its structure. If we consider the different question of what an observer can learn from looking at the resulting structure, such as the tree rings, or the fossil record, we are no longer concerned with information channels. For now, we can recapitulate that information transfer does not require pre-assigned meaning of two coupled events, such as the words in a language, and that the coupling must be unidirectionally unequivocal (the signal at the receiver must be coupled with only one signal from the source). The absence of equivocation guarantees the truthfulness of the informational content. If we are supposed to learn about a contingent matter of fact, we must receive truthful information about this in order to know. A contingent fact is one that could have turned out differently as far as the laws of nature or laws of logic are concerned. The current whereabouts of my neighbour's cat is an example of such a contingency. I cannot possibly derive it from anything else I know (if I haven't just seen the cat, and assuming I haven't had the chance to study the behavioural patterns specific to this animal), so in order to know I have to receive this information via a (Shannon) channel. We need truthful information because we have no (or few) means of deriving contingent matters of fact from some prior knowledge. If the justification whereby we account for knowing something is decoupled from a channel that has the matter in question as its source, truth and justification can become desynchronized (Floridi, 2004a). Equivocation-free transfer of information therefore prepares the way for the extraction of the content from the signal.

3.5 Digitalization and semanticisation

Following Dretske's account of naturalization of information further, we must now become acquainted with his concept of "digitalization":

> A signal (structure, event, state) S carries the information that s is F in digital form if and only if the signal carries no additional information about s, no information that is not already nested in s's being F. If the signal does carry additional information about s, information that is not nested in s's being F, then the signal carries this information in analogue form.
> (Dretske, 1981, p. 137.)

"Nested information" is to be understood as follows: the information that t is G is nested in s's being F = s's being F carries the information that t is G.

A general example of the analogue/digital distinction is that of a picture that contains, among others, the representation of a woman. The statement "the picture contains the image of a woman" is the digital representation of this very fact, whereas the complete picture carries this fact in analogue form, since other pieces of information can be derived from the picture that are not nested in the proposition that the picture contains the image of a woman, such as information about her height and hair colour. Notice that the proposition, the digital representation, can carry some other facts as nested information, e.g. the fact that the picture shows the image of a human being, since all women are human beings. The more conventional reading of "digitalization" that you may be familiar with involves the discretisation of continuous numbers. This is actually covered by the above definition: if we discretise an interval of real numbers to an integer that represents the interval's median, the median does not carry the information about which of the numbers within the interval has been discretised. However, the real number does carry the information about which median it belongs to, therefore this latter piece of information is nested within the real number, and the real number is an analogue representation of the median. For

example, we consider a list of real numbers ranging from 0 to 5, say 0.1, 1.7, 2.2, 2.7, and 4.1. All numbers are rounded to their closest integer. Then the numbers 1.7 and 2.2 are rounded to 2, and they are, in the given context, analogue representations of the integer 2, but not vice versa.

During digitalisation information is irreversibly destroyed. If Alice writes down "the picture contains the image of a woman", but destroys the actual picture, we lose information. When Bob looks at what Alice has written, he cannot reconstruct the rest of the information in the picture. In this sense a digitalisation is similar to a deduction. Although Alice might learn something by deducing a conclusion from the set of premises, the conclusion does not enable Bob to infer back to the premises from Alice's conclusion, since the proposition at the end of the deductive chain is generally weaker than the propositions of the premises. The process is irreversible, just like the digitalisation of information.

The next relevant concepts are the three orders of intentionality. These Dretske defines as follows:

> 1st order of intentionality:
>
> All Fs are G as a matter of fact
>
> S has the content that t is F
>
> S does not have the content that t is G
>
> (Dretske, 1981, p. 172.)

For example, suppose it happens to be the case that all cars on the road that Alice is currently driving along are hybrid vehicles. But it does not follow that a signal S that carries the information that vehicle A is on the road also carries the information that A is a hybrid vehicle.

> 2nd order of intentionality:
>
> All Fs are G according to natural law
>
> S has the content that t is F
>
> S does not have the content that t is G
>
> (Dretske, 1981, p. 173.)

All water expands when it freezes. If Alice knows this, she knows that when water is freezing in front of her, it is expanding. But suppose Bob does not know this fact, and his belief state S has the content "the water in this glass is freezing". His belief does not necessarily also have the content "the water in this glass is expanding".

> 3rd order of intentionality:
>
> All Fs are G according to analytic necessity
>
> S has the content that t is F
>
> S does not have the content that t is G
>
> (Dretske, 1981, p. 173.)

Alice might be able to tell whether a geometrical figure is a square; however, she might not be able to account for this judgement by stating that the figure has four sides of equal length, even though the latter would follow by analytic necessity from the fact that the figure is a square.

The second order of intentionality is not relevant to our present considerations, but the difference between the first and the third order is important. Initially, let's look at the first order to understand the importance for information processing systems in general. We mentioned earlier that correlated systems are not necessarily informationally coupled systems. In order to clarify this difference, consider the following situation: Alice and Bob work for two different companies. Carol implemented a scheme at Alice's company that determines when exactly workers are allowed to take a break from work and go to the canteen to have lunch. This is signalled by a little pop-up window in each employee's electronic calendar. Carol previously worked for Bob's company and implemented the same scheme there. So we define event A as "noon break starts for Alice" and the informationally coupled event B as "electronic signal pops up on screen". Events C and D are defined correspondingly for Bob. So, whenever D happens and Bob is thereby signalled that the noon break has started at his place, he can also tell that noon break has started for Alice. But this conclusion hinges on the coincidental fact that it was Carol who was responsible for the policy at both companies. The correlation between D and A is a spurious one and not genuinely informational. Contrast this with a situation where Bob has managed to wiretap event A, the beginning of the noon break at Alice's office, because he wants to have a little chat with Alice each time they both go to the canteen. Bob synchronizes event A with an event in his electronic calendar, let's call this event D*. D* now informs Bob about A via an information channel, and it has acquired an intentional state of the first order with respect to A. We can confirm this by applying the criteria above: D* has as content A, but not B. The signal D* does not tell Bob how the employees of Alice's company are informed about the noon break, since this is not part of its informational content, although B and D* are perfectly synchronized. In Dretske's words:

> To describe a physical state as carrying information about a source is to describe it as occupying a certain intentional state relative to that source. If structure S carries the information that t is F, it does not necessarily carry the information that t is G even though nothing is F that is not also G. The information embodied in a structure defines a propositional content with intentional characteristics. (Dretske, 1981, p. 172.)

Now, an intentional state is not sufficient to form a semantic structure, i.e. one that is capable of forming a belief state. We just saw that an intentional state with the first order of intentionality is just enough to distinguish informationally coupled systems from systems that are coincidentally correlated. Opposed to both of these, a genuine semantic structure must be capable of having as its propositional content one that is subject to the third order of intentionally: 'Any propositional content exhibiting the third order of intentionality is a semantic content'. (Dretske, 1981, p. 173.)

One can believe that one is shipwrecked in the Pacific Ocean without believing that one is shipwrecked somewhere between the western coast of America and the eastern coast of Asia and Australia. This is essentially Dretske's reading of a belief state. By means of the equivocation-free channel and the intentional states, we already understand better how *reference* is possible, how one structure can carry information *about another* structure. For example, the experience of a sound we hear is intentional with respect to the object that produces the sound, not with respect to the vibration of our ear drums, although the latter is implied by our having the experience. Now, with regards to how natural information can give rise to *symbolic* communication, another important idea is required: the difference between a piece

of information about a particular matter of fact, and the corresponding content the information has on a type level. Propositions of the form "s is F" require the concept of F. Knowledge of concepts is a form of type-level knowledge, and that is what is required for symbolic communication. The evolution of a symbolic system is connected to the problem of explaining how *false beliefs* can arise. If an individual s is falsely predicated by F, the concept of F, which has been acquired on informational (thus truthful) grounds relative to other individuals, is applied to the individual s without there being any information tokens, i.e. pieces of information belonging to a specific situation, that sustain the corresponding judgement. A false belief has arisen with respect to s. A meaning is a symbol connected to a concept on type level. The symbol can represent a concept in the sense of a genuine convention, but this is not necessarily so, since a symbol can gradually acquire its meaning.

The reader might still wonder why certain systems like digital measuring devices are not capable of carrying semantic content. Dretske discusses the example of a digital voltmeter whose gauge shows discrete numbers. This information, according to Dretske, is not "completely digitalised". The informational content embedded in the display carries information about what brought the displayed number about, via nested causal information between its inner components. By contrast, a genuine belief carries only the information (if truthful) that it bears in virtue of its meaning, the exact circumstances of how the belief was brought about are not accounted for, since the latter would concern different beliefs that are opaque against the belief's content. The fact that we have to, as it were, "forget" these circumstances, represents another aspect of the information loss that is crucial to the process of semanticisation.

3.6 Other approaches

Dretske solves the problem of explaining reference by means of his *intentional states*. Although several events, all part of a longer causal chain, precede the reception of a signal, the signal carries information about one primary event for the receiver. Radu Bogdan follows a similar strategy to Dretske, but bases the semantics of information more explicitly on teleology. We will outline his approach, before having a look at the more sceptical result from Quine's analysis. While pursuing his project of naturalization of epistemology in general, he came to the conclusion that meaning cannot be given an ontological status on a par with other real objects like the objects of physics.

3.6a Radu Bogdan's teleological approach

One can interpret Dretske's semantics in such a way that they are teleological, i.e. information is semantic if it is relevant to a *goal* that the receiver of the information has. Ragu Bogdan (1988) has an account of semantic information that is *outright* teleological. He starts with considering "material information", which is constituted by physical structures that causally influence each other in such a way that a certain quantity, which can later be identified with information, is preserved. For the sake of our overview, the concept of material information can be identified with evidential information. In both cases, at this stage we are still considering the flow of information, not its interpretation or utilization. In an attempt to give a naturalized account of the interpretation and utilization, we first encounter, again, the problem of reference. Material information is a causal concept, so how does an event at the receiving structure refer to one of the several events from preceding the causal chain?

Bogdan distinguishes "vital goals" and "active goals" of life forms. A vital goal would be the consumption of an apple, whilst the active goal would be to find out where the apple is situated, such that it can be grasped in order

to be consumed. The reference, the selection of which of the preceding causes of a signal is relevant, comes from the role the active goal plays for the vital goal. The information that is the satisfaction of the active goal is not yet the satisfaction of the vital need of the life form, but it enables the selection of the relevant cause: it is the whereabouts of the object, not the light that transports the information to the eye, which is significant. Such a goal-directed alignment of input and output is needed if we want to have a semantic artefact. A thermostat processes information (ambient temperature) and makes use of the information by turning on or off a radiator, such that it controls the ambient temperature. Still, it does not count as a semantically enabled artefact. The stimulus-reaction scheme (basically a causal scheme) of the thermostat must be extended by an intermediate step that semantically encodes the input. By this intermediate step, which comprises what Bogdan calls internal selectivity and intentional alignment, a fact that is relevant to the goal of the semantic system becomes active and triggers behaviour that leads to achieving the goal.

3.6b Quine's naturalization of meaning

Quine is famous for his project of naturalising epistemology, an endeavour that seemed necessary to Quine after he thought he had shown that no predication of a subject term follows analytically and by necessity from the meaning of the term (Quine, 1960). This hinges on an interpretation of meaning as an indefinite term. Here we will briefly summarize how he arrived at such a view. Quine demonstrated the ideas of a naturalised semantics by means of a thought experiment he called "radical translation". (Quine, 1960, p. 26.) Radical translation is an in-the-field technique of acquiring a foreign language when the translator has no prior knowledge of the other language. Instead, the translator must try to build up his grasp of the language by exposing himself to situations shared with a native speaker and by prompting assent or dissent to verbal utterances that are hopefully connected to the shared situation. Since assent or dissent refers to whole sentences uttered, the meaning of terms that are part of a sentence can be revealed only by an analytical process that is somewhat speculative and has no definitive endpoint. The speculative outcomes of such an activity Quine called "analytic hypotheses". He chose this model of language acquisition because he asserted that it is – to all intents and purposes – a faithful reconstruction of a child's learning of a language, and therefore how meaning arises with respect to terms handled by native speakers. Also, an inter-subjectively observable situation is a good starting point for an epistemology if its point is to come up with an ontology of things that really exist, as opposed to what is merely fancied to be true by single persons.

Unfortunately for meaning, the outcome of Quine's analysis did not admit it in such an ontology of real things. Meaning cannot be fully naturalised according to Quine – unlike the physical representation of what bears the meaning, i.e. the symbol. That is because the translation manuals that two different in-the-field translators come up with can be very different from each other, if they have worked independently of each other. Fundamental differences in meaning of sentences can arise even if the predication by truth values of two possible translations matches the truth value of a native sentence, or, in other words, two different translations of a native sentence of assent with respect to a situation would likewise each prompt assent within the community of native speakers of the translated language, although the two translated sentences differ in meaning. For Quine, things must be identifiable in time in order to grant them the status of existing things. Since meanings are indefinite and, a fortiori, not identifiable, they cannot be granted such a status.

What about the other aforementioned semantic structures, the bearers of intentional states (in particular, belief)? We can best compare their treatment by Quine if we look at his analysis of what Bertrand Russell called "propositional attitudes". If Alice believes that the evening star is a planet, but is unaware of the identity of the evening and morning star (see Chapter 6), she might fail to acknowledge that the morning star is a planet.

Frege's treatment of the propositional attitude of belief was to consider it a binary relation of a believer and the content of the belief, something generally called a proposition, which is taken to be independent of the language in which the proposition is expressed. Treating the proposition as a whole entity forbids the substitution of "morning star" for "evening star" in the aforementioned belief of Alice's, an operation that would render the statement untrue. Quine, however, cannot resort to Frege's solution of the problem because of lacking a criterion of identity of propositions. Propositions are allegedly identical if they have the same meaning, but meanings are, according to Quine, themselves not subject to such a criterion. Instead of using the notion of a proposition, his treatment of belief contents resorts to direct quotation: "Alice believes that the evening star is a planet" becomes "Alice believes 'the evening star is a planet'". "The evening star is a planet" is taken as a sentence, and sentences in turn are considered linguistic forms – it is not their actual utterance that matters but their belonging to linguistic forms, i.e. the classes[6] of utterances that would all have the same extension and would therefore all prompt assent if expressed in a language that the one whose assent is prompted understands. As opposed to Dretske's informational account, Quine's negative result in naturalising propositional attitudes therefore questions the reality of intentional states.

3.7 Summary

The process of becoming informed can be decomposed into the flow or transfer of information via a Shannon channel, and the semanticisation of the information. The first can be described on the basis of environmental information, whereas the second is not an instance of information transfer at all, since it involves information loss that contradicts the absence of equivocation, a necessary criterion for the transfer of information. To explain what happens during the semanticisation of information is a much more controversial aspect than the flow of information, and there is some doubt that the former can be naturalized at all.

3.9 Exercises

1. Revisit the poker game example of environmental information in human behaviour. Now assume two players work together against the other participants of a card game (not necessarily poker) by exchanging information about the cards in their hands between each other. Before the game, they agree on faking typical gestures of nervousness to signal their hand, picking a unique gesture for every card considered. Why is this not an example of environmental information?

2. How is environmental information related to the question of truth?

3. How is environmental information related to instructional information (see Chapter 13)?

3.10 Suggestions for the exercises

1. Environmental information is a natural coupling of two events. In contrast, the communication between the two players is a gestural language that can only work if the semantics of the gestures is agreed upon before the game starts. Every gesture encodes a specific card, therefore semantic information is transferred. Unlike this language by gestures, the natural reaction of players when looking at their hands is not semantic. Although it is true that the causal connection between the hand and the reaction is established by the

[6] Quine's naturalistic ontology admits physical objects as real entities, but also abstract objects like classes, if positing them simplifies our overall scientific theory of the world.

semantics of the cards (what they mean in the context of the card game), it is not the case that the reaction is itself a symbolic representation, i.e. a code, of the cards that the player holds.

2. Truth is not a property applicable to environmental information because it is a naturally established link between two systems, rather than a symbolic representation. Therefore, environmental information cannot misrepresent a state of affairs. The event observed at the receiver is dependent only on the mechanism of the information channel and the way it is measured (the level of abstraction).

3. The example of genetic information shows that instructional information is not necessarily semantic (symbolic) information. We must bear in mind, however, that it depends on our interpretation to see a realised "instruction" in a ribosome synthesizing exactly the proteins according to the genetic code. This interpretation, in turn, depends on our notion of a "natural purpose" that is fulfilled by the protein synthesis.

3.11 Further reading

A detailed description of the relevance of Shannon information to the flow of information, as well as explanations of the concepts of digitalization and a discussion of concept formation and adequacy of behaviour, is given in Dretske (1981).

A newer collection of essays by Dretske continues his programme of naturalizing the mind, with a focus on belief state formation Dretske (2000).

Quine's naturalization of epistemology is described in (Quine, 1960)

Part II: Social and moral

Chapter

4

4. ETHICS

The ethical implications of information

4.1 Introduction: Ethics and information

Ethics (or moral philosophy) is that branch of philosophy that investigates the concepts of right and wrong, both in themselves and in relation to human behaviour. Ethics is subdivided into sub-fields, the three most influential being metaethics, normative ethics and applied ethics.

The literature on ethics is vast and articulated and even a brief review would be well beyond the scope of this chapter. Therefore, instead of focusing on a definition of ethics as given by an author or in the context of a specific ethical theory, here the reader is presented with some general concepts to support the analysis offered in the next sections of this chapter.

Ethics (or moral philosophy) is that branch of philosophy that investigates the concepts of right and wrong, good and evil, both in themselves and in relation to human behaviour. Ethics is subdivided into sub-fields, the three most influential being metaethics, normative ethics and applied ethics. Meta-ethics is concerned with the nature of ethical concepts, judgements, propositions and dispositions. Typical meta-ethical questions are: "What do we mean by right and wrong?", "Are there ethical propositions?", "Do ethical judgements have a universal or relative scope?", and "What is the relationship between right and wrong, feelings and reason?"

Normative ethics focuses on ethical behaviour and on the criteria that should be adopted in order to behave ethically. In this context, several ethical frameworks have been devised, some of the most well-known being virtue ethics, deontology and consequentialism. Each normative framework defines a different way of identifying ethical behaviour. Virtue ethics defines good behaviour as one that promotes virtues (e.g. wisdom, courage, temperance) and avoids vices (e.g. cowardice, insensibility, and injustice). Deontology focuses instead on principles that have to be respected with no exception in order for one to be right. Depending on the type of deontological theory adopted, duties vary from being healthy to the Kantian moral imperative.[7] Finally, consequentialism discriminates between ethical and unethical behaviour by looking at the consequences of one's action. When the positive consequences outweigh the negative, an action is ethical; otherwise, it is

[7] A moral imperative in Kant is a *categorical imperative*, a requirement for every human being to act only in accordance with those principles that would always be valid for every rational human being. In Kant's words: "Act only according to that maxim whereby you can, at the same time, will that it should become a universal law." (Kant, 2002).

unethical. As with deontology, there are several types of consequentialist theories, depending on the type of consequences or on the recipients of the actions (patients) that should be considered.[8]

Now that we know that ethics and specifically normative ethics is concerned with right or wrong and our behaviour, it becomes relevant to ask whether there is a relationship between right or wrong, behaviour and information. Think, for example, about buying groceries at the supermarket. In the UK, often one of the first questions asked by the cashier, either human or mechanical, is whether the customer has a loyalty card. But what are the implications of using a loyalty card? Is recording information about when, where and how often every single item is bought ethically neutral? Is there a problem with customer privacy? What about whether and how the customer data are shared with other companies? Is it right or wrong for the supermarket to share customer data? Analogously, is posting personal information on one's Facebook profile morally neutral? What about sharing and tagging pictures online? Information is produced – commenting on an event, taking a picture – and publicly shared. Does the individual to whom the shared information relates – what we call the 'patient' of that sharing action – have the right to be asked for permission for such a sharing? The very same question applies to the paparazzi and the celebrities that they target. What about spamming or hacking? Is there something like a right not to be spammed? Is spamming right or wrong? Should the defacement of a website or the breaking of an electronic mailbox be considered immoral – aside from being illegal? What about sharing information about a company's balance sheets or, more subtly, accessing metadata[9] about electronic communications?

All these questions pertain to several domains but all share a similar characteristic: they all involve information and whether such information plays or should play a moral role. Specifically, the problem underlying all these questions is how one should behave when information is, for example, collected, shared, produced or destroyed. This is the general question to which Information Ethics shall give an answer. Looking at the different fields of ethics, it is clear that information ethics poses questions of normative ethics and metaethics: the issues are about how one should behave (normative ethics) and how right and wrong should be defined when information and its dynamics are considered (metaethics). Looking at the questions posed above, it is also clear that information ethics might span multiple domains and then be part of several branches of applied ethics – computer ethics, medical ethics, business ethics, librarian ethics and so on.

The following sections of this chapter offer an overview of how a unified approach to information ethics has emerged from research in multiple areas of applied ethics. The unified approach proposed by Luciano Floridi, the current mainstream information ethics theory, will be described, outlining its key characteristics and challenges. A section will be then dedicated to some of the most common misconceptions involving information ethics. Exercises and suggestions for further reading will close the chapter.

[8] For a general introduction to ethics see Baggini and Fosl (2007) or Blackburn (2002).

[9] The literal meaning of the word 'metadata' is data about data. In this context, it refers to data about a communication as opposed to the content of the communication itself. For example, the metadata about a phone call could include time, location, duration, caller and receiver identification numbers of that call.

4.2 Historical overview

Information ethics is a relatively young research field that is enjoying a rapid and prolific expansion. In order to understand the theories that are being developed today in this field, it is necessary to look at its history and at how it has evolved from multiple disciplines.

Timeline

1940-1950	*Cybernetics and first notes about the ethical implications of technologies for automation.*
1980-1990	*Computer Ethics and focus on the ethical issues associated with computers and, later on, with networking technologies.*
1990-2000	*Information Ethics fragmented and developed in the context of other ethical frameworks.*
2000-2010	*Information Ethics grows into an independent discipline.*

The origins of information ethics can be traced back to the 1940s and 1950s and are intertwined with the development of computer science and the information and communication technologies (ICTs). Information ethics began with the work of Wiener, a mathematician and professor at the Massachusetts Institute of Technology (MIT) who, for the first time, pointed out the social and ethical implications of developing electro-mechanical devices. Wiener fathered what he called "cybernetics" (Wiener, 1948), an interdisciplinary research field dedicated to the study of information feedback both in natural and artificial systems (see Chapter 1). While studying and envisioning the creation of machines capable of autonomy and information processing, Wiener developed a clear and mature understanding of the potential impact of such technologies on humankind and society. As a consequence, he was among the first to stress the need for specific ethical principles to guide the design and the deployment of such new systems. In his words, 'Long before Nagasaki and the public awareness of the atomic bomb, it had occurred to me that we were here in the presence of another social potentiality of unheard-of importance for good and for evil' (Wiener, 1948, pp. 27-28).

The ethical concern raised by Wiener remained marginal and largely ignored for more than a decade, before being brought to the fore again between the 1960s and 1970s when "Computer Ethics" emerged as a new field of study. Questions concerning computer-based crime (i.e. cyber-crime), disasters caused by computer errors or failures, or breach of privacy by accessing computer databases started to be noticed, and perceived as ethical issues. Computer ethics originated as a type of professional ethics, which inherited from cybernetics the concern for establishing ethical guidance (i.e. ethical policies) for the design and use of computers. For this purpose Parker wrote Rules of Ethics of Information Processing (Parker, 1968), which collected a set of examples of illegal and unethical uses of computers, providing the ground for the Code of Professional Conduct for the Association of Computing Machinery (Council, 1992). At the time, computers were providing new professional roles and offering new ways to create and manage information. There was a need for these new activities to be regulated, both legally and ethically.

The scenario changed between the 1980s and 1990s. In those twenty years, what we know today as the information revolution had begun. Personal computers were becoming ubiquitous, and the internet was starting to develop into a new global space for information sharing and commerce. The "vision" that computers would change the way in which people interacted with each other and with the environment became a reality, and in the span of just twenty years, individuals not only learned about, but became more and more dependent on, networked computing technologies to communicate, buy and sell goods, work and entertain themselves. This

was a major change in the ordinary lives of men and women and it also engendered a major change in the way the ethical problems related to the use of computers were regarded by computer ethicists.

In the middle of the 1980s, while some scholars still regarded the problems posed by the dissemination of computers as '[...] new versions of standard moral problems and moral dilemmas, exacerbating the old problems, and forcing us to apply ordinary moral norms in uncharted realms' (Johnson, 1985), others started pointing out that the problems related to the uses of computers were not simply a new version of old ethical conundrums but radically new problems engendered by the specific nature of computers and ICTs.

In this respect, the analysis proposed by Moor (1985) was ground-breaking. Moor stressed that there was something truly revolutionary about the social role of computers. According to Moor, computing machines were universal tools, for they could potentially perform any operation as long as the operation was defined in terms of input, output and logical operators. Moor stressed that after an initial phase of "technological introduction" (that is, the development and refinement of computing technologies), a phase of permeation would have followed in which we would have witnessed the capillary diffusion of computing technology in everyday life, thanks to the ability of computers to perform a vast number of useful operations. According to Moor, this spread would have induced radical changes in the way we worked, interacted and considered fundamental concepts like money or education. Such radical changes unveiled, for the first time in the history of computer ethics, the need to provide a tailored conceptual ground for defining the ethical principles to guide the use of computing technologies.

Moor, along with scholars working in the field of computer ethics, indicated that the dissemination of computers generated a "policy vacuum". There was a gap in the existing laws and policies of western countries, which were unprepared for regulating the design and use of such new technologies. By the end of the 1980s, it had become clear that the policy vacuum rested on a conceptual muddle, and that filling the vacuum required first a clear understanding of the nature of the ethical problems posed by the use of computers. In Moor's words:

> One difficulty is that along with the policy vacuum there is often a conceptual vacuum. Although a problem in Computer Ethics may seem clear initially, a little reflection reveals a conceptual muddle. What it needed in such cases is an analysis that provides a coherent conceptual framework within which to formulate a policy for action.
> (Moor, 1985, p. 269)

By the 1990s, it was clear that Moor, and many others, were right. The ethical problems engendered by the use of ICTs were not simply old problems in new shoes. Computers and computing technologies were starting to change the way individuals acted in the world, affecting social, political and economic infrastructures. It had also become clear that computing technologies raised new ethical problems, which required new theoretical approaches to define ethically-sound policies. Information was at the core of the ethical issues created by computers. Computing technologies were a new, powerful medium that pushed information processing to an unprecedented scale to such an extent that the ethical dimension of information impacted society and individuals with unprecedented strength. The "greasiness" (Moor, 1985, p. 269) of computing technology combined with the power of information was finally recognised and Information Ethics was born.

4.3 Towards a unified approach to information ethics

The previous section showed how the ubiquitous use of information technologies and the growth of the internet contributed to exposing the importance of the ethical dimensions of information. Following this impulse, information ethics developed across multiple research areas, often as part of ethical theories concerned with specific domains as, for example, medical, business, professional or librarian sciences. Altogether, the contributions coming from these fields can be classified as information ethics but the lack of a unified approach and, to some extent, of a metaethics grounded on purely informational principles hindered further developments. This section offers a review of different approaches to information ethics and an analysis of their strengths and limitations.

Floridi (2010a) distinguishes different approaches to information ethics depending on three ways of considering information: information as resource, information as product and information as target.[10] Information can be seen as a resource when an individual acquires and possesses information, a process that is often relevant from an ethical point of view. For example, in medical ethics informed consent is possible only in the presence of complete and understandable information. Analogously, accurate diagnoses are possible only when access to patients' information is available. Conversely, acquiring and possessing information is not always a means of behaving morally; sometimes it is actually quite the opposite. In professional or computer ethics, respecting anonymity – so avoiding acquiring information about someone's identity – is often necessary to guarantee a fair and impartial treatment of individuals. Think about the importance for students of submitting their essays anonymously in order to be marked fairly.[11]

Information is a product when it is elaborated, managed and shared by agents such as individuals, companies or institutions. Also in this case, the act of producing information may have ethical consequences and is a subject for ethical enquiry. Again, the quality and quantity of the information produced may be used to define whether a single individual, a group of people, companies or even states are behaving ethically. So, for example, in business ethics companies are considered to behave ethically – and legally – when they make available untainted balance sheets to the public and to the rating agencies. In professional ethics, students behave ethically – and legally – when they hand in original essays instead of cut and pasted patchworks obtained from the internet. Politicians and citizens debate whether states behave ethically when promoting free and independent press agencies instead of fostering propaganda.

Information becomes a target when we consider an environment that is made of information. Informational environments are more common than might be initially thought. Consider, for example, the operating system of a computer. It consists of a set of programs, some useful to drive the physical components of the computer, others necessary to interact with the user or to communicate across the internet. Applications allow for the creation of documents, pictures, or movies, all stored in the storage unit alongside the other components of the operating system. Altogether, these applications and files constitute an informational environment. Even better, an informational environment does not need to be digital. A library, an archive, or a land registry are all examples of analogue informational environments. How individuals should behave in such environments or how these environments should be preserved and managed is a matter of ethical investigation. In computer and environmental ethics, for example, hacking, piracy, filtering, vandalism, or archive preservation are all ethical issues that stem from considering information as a type of target.

[10] The rest of this section and the following one are an adaptation of the material published in (Floridi, 2010a). Permission from the author has been granted.

[11] Note that from a metaethical point of view, the quantity and quality of available information is directly linked to the concept of moral responsibility. After all, poorly informed individuals might behave unethically even with the best of intentions.

Approaching information ethics by means of these three ways of defining information is useful. It allows many fields to recognise the ethical relevance of information – see medical, professional, business, environmental, and computer ethics in the examples above – and clearly shows the important role played by technology when considering the moral sphere of information. Nonetheless, grounding information ethics only on consideration of information as a resource, as a product or as a target, and fragmenting it across multiple research areas leads to two main limitations. On the one hand, this approach is still too simplistic. Arguably, several important ethical issues belong mainly, but not only, to the analysis of just one way of considering information. On the other hand, the approach is insufficiently inclusive. There are many important ethical issues that cannot easily be placed on the map at all, for they really emerge from, or supervene upon, the interactions that go on through acquiring, producing or targeting information (see examples below).

As soon as we consider ethical concepts that involve articulated dynamics of information, the simplistic nature of an ethics that focuses only on one way of considering information becomes clear. Consider, for example, censorship, misinformation and free speech. Censorship affects an individual both as a user and as a producer of information by forbidding not only the publication but also the acquisition of information. Misinformation (i.e. the deliberate production and distribution of misleading, false content) is an ethical problem that concerns all of the three ways of considering information. Freedom of speech affects not only the production of information but also the availability of offensive content (e.g. violent content and socially, politically or religiously disrespectful statements) that might be morally questionable and should not circulate.

Coming to the lack of inclusiveness, ethical issues involving some properties of information that are not connected to the dynamics or flow of information cannot be addressed by ethical theories that consider information only as a resource, a product or a target. These theories prescribe whether some individual should receive or share some information, what characteristics such information should have, whether and how holding information relates to the individual duties or to the consequences of the individual's behaviour. If the information does not flow, is neither received nor shared, then these theories can say little about how the individual should behave or what ethical principles should apply in a given situation. So, for example, when the ownership of some information is considered, the relevant problem becomes whether that information has been created by an individual not whether that information should be received or shared.

The information ethics proposed by Floridi aims at addressing these limitations. It does so by endorsing an inclusive definition of information, by acknowledging the moral relevance of the dynamics of information and by introducing a radical, universal notion of an informational entity. In this way, Floridi proposes a unified and independent theory of information ethics that extends well beyond the limitations in scope and inclusiveness suffered by those theories that endorse a single way of looking at information.

4.4 Floridi's information ethics

Floridi's information ethics is built on three fundamental concepts: information ontology, the agent/patient pair and the infosphere. These concepts need to be clarified before we come onto the details of Floridi's ethical theory.

Floridi's information ethics looks at information as an entity, thereby endorsing an ontological approach. Imagine looking at the whole universe from a chemical perspective. Every entity and process will satisfy a certain chemical description. A human being, for example, will be between 45% and 75% water. Now consider an ontological informational perspective. The same human being will be described as a cluster of data, that is, as an informational entity. Please note the words "will be described as". They are important as they stress that the

entity is described in those terms, not that a human being – or any other entity – is only or essentially a cluster of data. It is just a way of looking at entities or, more correctly, the explicit choice of a Level of Abstraction (see Chapter 2).

The agent/patient pair stands for any informational entity that either produces some effects on the environment (an agent) or is the recipient of such a change (the patient). The ontological approach just described implies a very inclusive definition of an entity. An informational entity does not need to be alive, let alone conscious or even embodied. Therefore, at a given level of abstraction, an informational entity, either as an agent or as a patient, can be a person, animal, and plant but also anything that exists, from a painting and a book to a star and a stone; anything that may or will exist, like a future generation; and anything that was but is no more, like one of our ancestors or an old civilization, or even an ideal, intangible or intellectual object. From this perspective, informational systems, rather than just living systems in general, are raised to the role of agents and patients of any morally relevant action, with environmental processes, changes and interactions equally described informationally.

The infosphere is the sum of all the informational entities and of their relations. It can be thought of as the informational equivalent of the biosphere as long as we remember that at a given level of abstraction the biosphere can also be considered informationally.

We are now ready to look at Floridi's definition of information ethics (Floridi, 1999):

> Information Ethics is an ontocentric, patient-oriented, ecological macroethics

An intuitive way to unpack this definition is to compare information ethics to other environmental approaches. Biocentric ethics usually grounds its analysis of the moral standing of bio-entities and eco-systems on the intrinsic worthiness of life and the intrinsically negative value of suffering. It seeks to develop a patient-oriented ethics in which the patient may be not only a human being, but also any form of life. Any form of life is deemed to enjoy some essential moral rights that deserve and demand to be respected when contrasted to other interests. So biocentric ethics claims that the well-being of the living entities ought to contribute to guiding the agent's ethical decisions and constraining the agent's moral behaviour. In this context, patients are placed at the core of the ethical discourse, as a centre of moral concern, while agents are moved to its periphery.

Now substitute "life" with "existence" and it should become clear what information ethics amounts to. Information ethics is an ecological ethics that replaces biocentrism with ontocentrism and suggests that there is something even more elemental than life, namely being – that is, the existence and flourishing of all entities and their global environment – and something more fundamental than suffering, namely entropy. The latter is most emphatically not the physicists' concept of thermodynamic entropy. Entropy here refers to any kind of destruction or corruption of informational objects (mind, not of information), that is, any form of impoverishment of being, including nothingness, namely the not being. More specifically, destruction is to be understood as the complete annihilation of the object in question, which ceases to exist, while corruption stands for a form of pollution or depletion of some of the properties of the informational object.

Floridi's information ethics holds that being/information has an intrinsic worthiness. He substantiates this position by arguing that any informational entity has a right to persist in its own status, and a right to flourish, i.e. to improve and enrich its existence and essence. As a consequence of such "rights", information ethics should evaluate the duty of any moral agent in terms of contribution to the growth of the infosphere and any

process, action or event that negatively affects the whole infosphere – not just an informational entity – as an increase in its level of entropy and hence as an instance of evil (Floridi, 2003; Floridi & Sanders, 1999, 2001).

When so conceived, information ethics is impartial and universal because it holds that every entity, as an expression of informational being, has a dignity which deserves to be respected by every agent of the infosphere. This ontological equality principle means that any form of reality (any instance of information), simply by the very fact of being what it is, enjoys an initial equal right to exist and develop in a way that is appropriate to its nature.

The conscious recognition of the ontological equality principle presupposes a disinterested judgement of the moral situation from an objective perspective, i.e. a perspective which is as non-anthropocentric as possible. The application of the ontological equality principle is achieved whenever actions are impartial, universal and "caring".

The crucial importance of the radical change in ontological perspective cannot be overestimated. Bioethics and environmental ethics fail to achieve a level of complete impartiality, because they are still biased against what is inanimate, lifeless, intangible, or abstract (even land ethics is biased against technology and artefacts, for example). From their perspective, only what is intuitively alive deserves to be considered as a proper centre of moral claims, so a whole universe escapes their attention. This is precisely the fundamental limit overcome by Floridi's information ethics, which further lowers the minimal condition that needs to be satisfied, in order to qualify as a centre of moral concern, to the common factor shared by any entity, namely its informational state. Since any form of being is in any case also (but not only) a coherent body of information, to say that information ethics is infocentric is tantamount to interpreting it, correctly, as an ontocentric theory.

4.5 The fundamental principles of Floridi's information ethics

Floridi's information ethics determines what is morally right or wrong, what ought to be done, what the duties, the "*oughts*" and the "*ought nots*" of a moral agent are, by means of four basic moral laws:

1. entropy ought not to be caused in the infosphere (null law);
2. entropy ought to be prevented in the infosphere;
3. entropy ought to be removed from the infosphere; and
4. the flourishing of informational entities as well as of the whole infosphere ought to be promoted by preserving, cultivating and enriching their properties.

The moral question asked by information ethics is not "why should I care, in principle?" but, "what should be taken care of, in principle?" Approval or disapproval of an agent's decisions and actions should be based on how the latter affects the well-being of the infosphere and hence the informational entities involved. The duty of any moral agent should be evaluated in terms of contribution to the sustainable blooming of the infosphere, and any process, action or event that negatively affects the whole infosphere – not just an informational object – should be seen as an increase in its level of entropy and hence an instance of evil.

The four laws clarify, in very broad terms, what it means to live as a responsible and caring agent in the infosphere. The laws are listed in decreasing order of importance. Breaking rule number 3 is less depreciable than breaking rule number 2. Breaking rule number 0, the null law, is the worst an informational agent can do, so the blame is the highest. Accordingly, an action is unconditionally

commendable only if it never generates any entropy in the course of its implementation; and the best moral action is the one that succeeds in satisfying all four laws at the same time.

Most of the actions that we judge morally good do not satisfy such strict criteria, for they achieve only a balanced positive moral value; that is, although their performance causes a certain amount of entropy, we acknowledge that the infosphere is in a better state on the whole after their occurrence. It should be noted that a process that satisfies only the null law – the level of entropy in the infosphere remains unchanged after its occurrence – has no moral value, that is, it is morally irrelevant.

4.6 Common misconceptions and further analysis

Floridi's information ethics endorses a fairly innovative approach to the problem of the moral dimension of information and is still under development. An innovative nature and a degree of incompleteness are a very good recipe for misunderstandings and misconceptions. Here we highlight some of the most common mistakes in addressing information ethics so as to shed some light on the misconceptions and, in doing so, to clarify what may otherwise seem controversial aspects of Floridi's theory.

We will focus on misconceptions concerning:

- the informational nature of entities;
- the "overridable" nature of the informational rights; and
- the ethical status of spam and viruses.

Typically, two misunderstandings can happen when considering the informational nature of entities. The first is the identification of what Floridi calls informational entities with any other piece of well-formed and meaningful data such as news, emails, the *Britannica*, or Leibniz's *De Arte Combinatoria*. It should be remembered that Floridi's information ethics is grounded on an ontological premise, namely that given the right level of abstraction the entirety of reality can be seen informationally. It is not just about meaningful data – i.e. semantic information (see Chapters 3 and 7) – but also about everything that exists, that can be seen informationally and that can be affected by the action of another entity. So, yes, a newspaper is an informational entity and, as such, it is worthy of moral consideration; but so is also a future generation, your ancestor, your garden and Bob's *Ficus Panda* bonsai.

Bob's bonsai brings us to the second misconception that is often associated with the informational nature of entities, namely the idea that considering an entity informationally is akin to reducing it to just a cluster of data. To address this misconception, it is enough to note that even if at a proper level of abstraction a bonsai can indeed be seen as a cluster of data, this does not exclude that at a more inclusive level the bonsai can also be seen as a living entity. Let us consider yet another example to refresh our understanding of what a level of abstraction is. Suppose Alice has just turned eighteen and that she has saved enough money finally to buy a car to travel across Europe during her summer vacation. She buys a cheap second-hand car, which, naturally, requires some fixing. To Alice, the car is the symbol of her achievement, autonomy and freedom; to the mechanic, the car is just an object with some broken parts; to its maker, the car is a set of engineering blueprints. After considering this scenario, suppose someone asks you: "what is a car? Is it a symbol of autonomy, an object to be fixed or an engineering blueprint?" The correct answer would be: "it depends". Depending upon the level of abstraction embraced, a car may well be something to be fixed or a bunch of blueprints, but this does not deny the possibility of it also being a symbol of autonomy and freedom for Alice. The whole point of information ethics is that everything can be described as an informational entity because the

informational level of abstraction is the lowest common denominator among all entities.[12] As in the case of Alice's car, focusing on the informational perspective does not deny the possibility of adopting other perspectives, such as, for example, a more human-centred one.

The clarification of the role played by ontology in Floridi's information ethics is a good starting point to cast some light on the second misconception, the one about the overridable nature of the informational rights. The misconception goes as follows: if all entities share the same nature, all have informational rights as informational entities, and all are potential patients in the infosphere, then we have a moral impasse because it is not possible to determine the relative moral value among a piece of software, Leibniz's writings and the kids playing in the park. All informational entities are equal when considered from the perspective of information ethics. The impasse is overcome when considering that all entities have some initial minimal rights to exist and flourish and that such rights can be waived (overridden) depending on the effect – consequences – that they have on the other informational entities. An example should help to make this more evident. Every citizen born in a democratic country enjoys by default some rights, for example that of being free, as in not imprisoned. This right is shared among all citizens and lasts as long as the citizen behaves according to the laws approved by a democratically elected government. Any citizen who may be proved to have breached the law loses the right to be free and is incarcerated. In the same way, all entities come into existence enjoying the right to exist and flourish as long as they do not cause entropy in the infosphere.

We are finally ready to consider the last misconception, concerning entropic information entities such as spam and computer viruses. In this case, someone might mistakenly argue that because all informational entities share some minimal initial rights and spam and computer viruses are indeed informational entities, then according to Floridi's information ethics, destroying or corrupting spam and viruses would be unethical. It should be clear by now that according to information ethics all entities have some initial minimal rights but that such rights are lost if the entities increase the amount of entropy in the infosphere. When this happens, Floridi's information ethics prescribes that the entity's initial rights should be overridden and that the offender should be altered or destroyed in order to stop it from inflicting entropy on the infosphere. It should also be noted that bringing entropy into the infosphere can and should be balanced by looking at how the same entity promotes and fosters the flourishing of the infosphere. If the entity contributes more to the infosphere than it destroys, then its initial rights should not be overwritten. In the case of spam and viruses, the trade-off is clearly biased towards entropy and as such, it would be ethical to prevent them from further damaging the infosphere.

4.7 Exercises

1. What is the difference between metaethics, normative ethics and applied ethics?

2. How would you classify computer ethics?

3. What is the difference between cybernetics and computer ethics?

4. What is the difference between professional ethics and Moor's computer ethics?

[12] The careful reader may wonder about how such a claim should be justified. Floridi's information ethics is grounded on an informational approach to metaphysics. Departing from a Newtonian perspective in which reality is made of concrete, static entities that are "outside and inside" the perceiving/knowing subject, Floridi embraces a metaphysics where reality is everything that interacts with us, both concrete and abstract, material or immaterial. Such a reality is made of relationships, interactions, transitions or fields called "informational structures". For this reason, *being* in information ethics can always be reduced to its (minimal) informational nature. It is such an informational nature that has an intrinsic value, a value that can be overwritten depending on the effects that the interactions of informational structures have on the whole – i.e. the infosphere.

5. What is the difference between computer ethics and information ethics?

6. What are the three ways to consider information that underline many different approaches to information ethics?

7. What are the benefits and limitations of a unified approach to information ethics?

8. What are the three concepts on which Floridi's information ethics is grounded?

9. Do you agree with Floridi's position that every entity can be described at an informational level of abstraction?

10. Do you think Floridi's information ethics offers a way to categorise informational entities by order of moral relevance?

11. It has been written that information ethics is still under active development. What is missing, in your opinion?

12. What is the metaphysical assumption of Floridi's information ethics?

13. Offer three examples of how Floridi's information ethics could be applied to everyday situations.

4.9 Further reading

Baase (2012), Epstein (1996), Thomson (1999), Floridi (forthcoming-b).

Chapter 5

5. SOCIETY

How does information affect society?

5.1 Introduction

Imagine that Alice has the goal of finding the cheapest hotel near JFK airport in New York. She has different choices: (a) she personally calls all the hotels in the JFK area asking for prices, (b) she trusts her travel agent and refers to him to find the cheapest hotel, or (c) she trusts a search engine, she types the query "cheap hotel JFK" and goes for the cheapest hotel mentioned in the answers. Can we say that Alice trusts the travel agent in the same way she trusts the search engine? What does this example tell us about the reality in which we live?

The information revolution concerns the increase in information and communication technologies (ICTs). During the past two decades such technologies have developed fast and acquired a crucial role both in individuals' daily practices and in the social, political and economic processes of current societies. Consider, for example, how many of our social interactions and how much of the weekly working or studying schedule we manage through ICT-based devices, like smartphones, laptops or computers; or think about how many resources and infrastructures of our societies depend on ICTs. The information revolution is deemed to be the origin of radical changes concerning the way we interact with others, with the environment, and with the very structure of the reality in which we live. Such changes reshape current societies, so much so that they are now referred to as the information societies.

In the rest of this chapter we will first recall details of the information revolution. We will consider the historical relevance of this phenomenon and its social and philosophical implications. We will then turn our efforts to the ethical problems that such a revolution engenders, and we will analyse the problems of online trust (Turilli, Vaccaro, & Taddeo, 2010) and information warfare (Taddeo, 2011). The conclusion will allow us to pull together the threads of the analysis developed in this chapter.

5.2 The information revolution, history and society

To some extent the information revolution could be compared to a piece of music that starts quietly and builds up slowly, beat after beat, until it explodes in a blast. The information revolution started around 3000 BC with the Sumerian pictographs. From that moment on, the information revolution has accompanied the history of mankind. Among the milestones of the information revolution are

Gutenberg's invention of the printing press in 1455; the work of Augusta, Lady Byron, Countess of Lovelace and Babbage on the Analytic Engine in the early 1830s; the invention of the first telephone during the 1870s; Turing's work during World War II; the development of ARPANET by the US Department of Defense in the 1960s; the first versions of the UNIX Operative System in the late 1960s; and the progressive dissemination of personal computers, laptops and smartphones begun in the late 1970s and continuing until today.

More than an information revolution, it seems that humankind is experiencing a long-lasting turn concerning the development of technologies able to create, transmit and store information. At first glance there seems to be nothing revolutionary about this development, but a more attentive analysis and the consideration of two aspects allows us to understand the revolutionary nature of the informational trend. The two aspects are: (i) the extensive dissemination of ICTs, so-called "ubiquitous computing"; and (ii) the profound changes that such dissemination creates in our societies. Let us consider them in more detail.

We witness the dissemination of ICTs every day. You may think of personal usage of computers for working or entertainment purposes as an example of such dissemination, but this is only the tip of the iceberg. ICTs provide the ground for the economic and industrial growth of our societies; they constitute one of the fundamental tools for the progress of experimental science and provide the means for storing and managing historical, economic, and legal information. Slowly and ineluctably ICTs have grown to the point of becoming necessary for societies and individuals to live and prosper. They provide new modes for creating and managing information, which lead to new means of interaction with other individuals and with the environment. Consider, for example, the way in which we perceive distances and time nowadays, when we can talk and see someone on the other side of the world or when we can exchange documents in a matter of seconds, rather than in the days or months it used to take only a few decades ago. Not only do such changes affect our personal experience, they contribute to redesigning the very structure and rules of our societies as well.

A noteworthy analysis in this respect has been provided by Castells (2000). The proposed analysis highlights a networking logic as the distinctive characteristic of information societies. According to this analysis, ICTs facilitate the organisation of social interactions in the shape of (complex) networks, and such networks constitute the backbone of current social and economic processes. The networking logic leads to a set of social changes, e.g. decentralization within firms, remote-working and interactions, development of the virtual community, and globalization.

Social networks, as Castells describes them, can expand without limits by integrating new nodes and are much more flexible and plastic, as they are not organised in any institutional shape. Societies organised according to the networking logic are radically different from their predecessors and for this reason they experience a policy vacuum concerning the management of social, political and economic phenomena that are governed by the networking logic.

Philosophers and ethicists argue that the new policies and regulations required to fill such a vacuum need to be grounded on a clear understanding of the nature of the information revolution and on new ethical principles prescribing whose and which rights to preserve while ruling the information society. Let us continue our analysis by describing the nature of the information revolution (see also Chapter 2).

5.3 The information revolution as the fourth revolution

In a nutshell, the information revolution radically changed the way in which human beings perceive themselves, the universe, and how they interact with the rest of the world. In this sense the information revolution is a truly philosophical revolution.

A philosophical analysis of the information revolution has been provided by Floridi, who refers to the information revolution as the fourth revolution (Floridi, 2007b, 2009). The information revolution is the latest revolution in the history of western culture to bring to the fore radical changes in the way human beings understand themselves and their place in the universe. Following an analysis already proposed by Freud (1917), Floridi argues that three conceptual revolutions occurred following the works of Copernicus (1473–1543), Darwin (1809–1882) and Freud (1856–1939). Each of these revolutions repositioned humanity with respect to the universe. (See section 2.2.)

The Copernican revolution revealed that humanity inhabited one planet among many orbiting the sun; the Darwinian revolution showed that humanity is not at the centre or at the top of the biological kingdom; and the Freudian revolution showed that our mind is far from being transparent to itself. If considered in succession, these three revolutions show a trend: they progressively dismantle the anthropocentric understanding of the universe, displacing humanity from a special position.

For Floridi, the information revolution, which we might think of as the Turing revolution, for his refinement of the concepts of algorithm and computation with what came to be called a Turing machine (Petzold, 2008; Turing, 1936) is the next step. Turing also contributed the Turing test to the debate concerning the possibility of developing conscious and thinking machines in Artificial Intelligence (Turing, 1950). The Turing test is the latest step in the trend of dismantling the anthropocentric approach to the universe. The test introduced for the first time the idea that thinking and being conscious, which had been considered human prerogatives for millennia, could to some extent be attributed to non-human entities, like machines.

It is clear that the information revolution triggers profound changes both in the way we conduct our lives and in the way we perceive ourselves as human beings. Such changes pose important conceptual and ethical issues (see Chapter 4). We will describe the former in the next section, by considering how individuals and the environment in which they live are conceived after the information revolution. We will continue treatment of the latter in section 5.4.

5.3a Informational organism and informational environment

We now need to reintroduce a philosophical word in our analysis. This is re-ontologization (see section 2.2), which refers to a process of redefinition of the (ontological) properties of the existing entities and their environment. In the rest of this chapter we will use this word to refer to this process that follows the information revolution.

The dissemination of ICTs drives the re-ontologization process. The process occurs in two steps. The first is the creation of a new domain in which new entities exist and new modes of interactions with and among those entities are made possible. This is the non-physical domain, constituted by virtual or digital entities interacting both with other non-physical entities, such as two computer programs interacting with each other, and with physical entities, such as completely automated software interacting with its users.

The second step is the merging of this new domain with the "old" physical one. This step of the process is known as "ubiquitous computing" or "ambient intelligence"; it has been occurring for a while and is now close to completion. Consider, for example, the latest smartphones, which can wait for you walk into the supermarket to remind you of the need to buy milk and which can alert you that you need to go to a 10am meeting as soon as you step into your office building.

Entities of the non-physical domain are part of reality and exist to the same extent to which the entities of the physical domain exist. The two domains are conceived of as integrated: they are both part of the *infosphere*, i.e. of the environment in which we live and with which we interact (see Chapters 2 and 4). This becomes evident if we think, for example, about how personal information that we post on an online social network affects our life offline or how much time we spend online while waiting for the next train to arrive or the dentist to be ready to see us (see Chapter 15).

The re-ontologization concerns both human beings and the environment. With the four revolutions, human beings were dethroned from their central position in the universe. The idea of humanity as a superior and unique species was demolished by biological and scientific studies, and now the erosion of the anthropocentric approach allowed the unveiling of the informational nature of human beings, who discover their nature as connected informational organisms, or *infORGs*. In Floridi's words:

> [The] fourth revolution is the process of dislocation and reassessment of humanity's fundamental nature and role in the universe. We do not know whether we may be the only intelligent form of life. But we are now slowly accepting the idea that we might be informational organisms among many others, significantly but not dramatically different from natural entities and agents and smart, engineered artefacts.
> (Floridi, 2007b, p. 62)

Despite being a trendy word, "inforgs" does not stand for the characters of the next sci-fi bestseller, nor should inforgs be mistaken for cyborgs, i.e. human beings with implanted ICT devices or avatars or online alter-egos. Inforgs are human beings understood as informational entities (see Chapters 4 and 15), who are enjoying a fully connected life in the informational environment.

In respect of the environment, the re-ontologization occurs because a substantial part of the environment in which we live is shaped by the creation, management and utilisation of information, communication and computational resources.

The re-ontologization process is the core of the fourth revolution. It is the source of some of the most profound transformations and challenging problems that information societies are experiencing and will experience in the near future, as far as technology is concerned. Now that we have a clear grasp of the nature and of the conceptual effects of the information revolution, we will consider in more detail two of the ethical problems that it engenders.

5.4 Ethical problems of the information society

Many of the ethical problems of the information society are often mentioned in the media or in everyday conversation with friends and colleagues. Most of us have at least a superficial understanding of the issues concerning privacy, the digital divide, online trust, online identity, information warfare and information crime, to mention just a few. One of the reasons why these problems are so popular is because they concern our everyday activities, from the information we make available online on, say, Facebook, to the security we are guaranteed while using online banking.

The solutions to these problems hinge on both a policy-oriented and an ethical approach. In the rest of this chapter we will focus on the ethical approach and analyse two problems: online trust and information warfare. These two problems have been selected because they concern two important aspects of our information society: online trust affects personal interactions occurring in the informational environment, while information warfare constitutes a new mode for information societies to interact with each other as a whole.

5.4a Online trust

The problem of online trust arose during the last decade with the dissemination of Web 2.0 and the consequent development of online social interactions. The problem arose because trust is considered as a characterising aspect of social interactions: as internet users started to develop online social behaviours, some scholars wondered whether trust could also emerge in online contexts (see for example Nissenbaum (1998)). In this section, we will first consider a brief overview of the debate on this topic; then we will analyse what online trust is and what role it has in online interactions.

Trust is generally understood as a decision taken by an agent (the trustor) to rely on another party (the trustee) to perform a given action. This decision rests on the assessment of the trustee's trustworthiness. The trustor's decision implies some risk that the trustee will behave differently from expected, and hence betray the trust in her. In order not to take too high a risk of betrayal, the trustor usually seeks some guarantees of the trustee's behaviour, i.e. he assesses whether the trustee is or is not trustworthy (Gambetta, 1998; Luhmann, 1979).

The debate concerning online trust is polarised by two positions. One advocates that trust cannot emerge online, because online interactions do not satisfy two necessary conditions for the occurrence of trust (Nissenbaum, 1998): (i) the presence of a shared cultural and institutional background, and (ii) unequivocally assessing the trustee's physical identity. The other position gathers those who argue in favour of the presence of online trust (see for example Weckert (2005)) by showing that either (i) and (ii) are not necessary conditions for the emergence of trust, or that it is actually possible to satisfy (i) and (ii) in online interactions.

According to the detractors of online trust a fundamental aspect to consider in deciding whether or not to take the risk of trusting another individual is the kind of social norms, cultural and moral values characterising the social environment. Following this thesis, trust emerges only if the trustee shares the same set of norms and values with the trustor. Such a shared background provides the guarantee that the trustee will behave as she is expected to do and, consequently, lower the risk of betrayal. Two reasons are presented in support of this thesis. First, the shared background provides a set of parameters recognised

both by the trustor and the trustee for assessing what correct behaviour is; and second, the trustee feels a social pressure to behave according to the shared norms and values, which will prevent her betraying the trustor.

Several theses have been provided in defence of the occurrences of trust online (Taddeo, 2009). All of them agree in considering trustworthiness the ground on which trust rests: 'trustworthiness is the guarantee required by the trustor that the trustee will act as it is expected to do without any supervision', (Taddeo, 2010, pp. 246-247). The differences arise when considering what trust is and what its role is in online interactions.

For example, Taddeo argues that when trust is present, the trustor decides to delegate the performance of a given task to the other agent i.e. the trustee, and not to supervise the trustee's actions. According to the author, delegation and absence of supervision of the way the delegated action is performed are the main characteristics of the occurrence of trust. So if Alice trusts Bob to buy some milk, Alice delegates the task to buy some milk to Bob and will not supervise or check whether Bob is actually buying some milk, nor does she supervise the way in which Bob buys the milk. The same happens for online interactions, when we trust an online retailer to sell the advertised products or to post them in the estimated time.

According to this analysis, trust is not a relation, but a property of a relation; that is, something that qualifies (changes) the way the relation occurs. The previous example will help to clarify this. Alice and Bob are in a relation even if Alice does not trust Bob to buy the milk, but Alice still asks Bob to go and buy the milk. The only difference is that in the circumstances in which Alice trusts Bob, Alice will not supervise Bob while Bob is supposed to perform the delegated action (Figure 6).

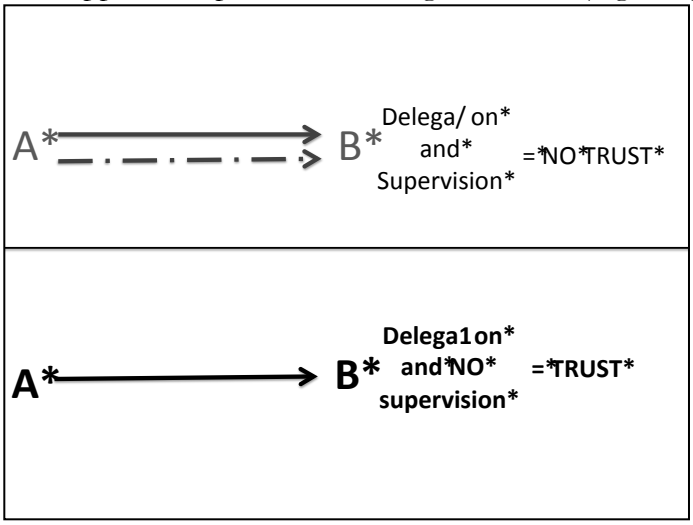

Figure 6: Trust as a property of a relation

This analysis unveils the fact that trust is advantageous for the trustor: when the trustor trusts the other agent she can delegate the performance of a given task without having to supervise the trustee's actions. In this way the trustor minimises her effort and commitment in achieving her own goal. This is true also for online occurrences of trust. Imagine, for example, Alice who aims to find the cheapest hotel near JFK airport in New York. She has different choices: (a) she trusts no-one and looks by herself for the telephone numbers of all the hotels in the JFK area and calls all of them asking for prices; (b) she trusts her travel agent and allows him to

find the cheapest hotel; or (c) she trusts a search engine i.e. she types the query "cheap hotel JFK" into the engine and goes for the cheapest hotel mentioned in the answers. Both (b) and (c) are relations characterised by trust and in both cases our agent has saved time and energy in achieving her goal.

This analysis of trust also allows us to understand the role of online trust. Online interactions are grounded on the communication of information (Corritore, Kracher, & Wiedenbeck, 2003), where "information" should be understood in the general sense of meaningful content that can be transmitted from a sender to a receiver (for the sake of simplicity in this chapter we will disregard the debate concerning the truthfulness of semantic information discussed in Chapter 7). So online trust occurs when in online relations of communication between Alice and Bob (recall the example of the agent looking for the cheapest hotel near JKF), the receiver of the information trusts the sender of the information (the search engine in our example) and therefore accepts the communicated information without checking its truthfulness and relevance or supervising the actions performed by the sender.

We can define online trust as a particular case of trust, characterised by two factors: it occurs in online interactions and it only qualifies relations of communication. Like trust, online trust is also grounded on the trustee's trustworthiness. Since online trust is successful when the communication between the sender and the receiver is honest and transparent, honesty and transparency are the criteria that should be endorsed in the assessment of the potential trustee's trustworthiness.

Online trust makes interactions advantageous for the trustor, hence providing an incentive for the trustor to interact with other individuals. As a consequence, online trust increases the interactions and the social network of the individuals in the system. Furthermore, individuals refine their social intelligence – i.e. they learn to evaluate the trustworthiness of the other correctly – to avoid risky interactions. This initiates a virtuous circle that leads to a selection process, according to which trustworthy individuals are involved in a growing number of interactions, whereas, in the long run, untrustworthy individuals are progressively marginalised and excluded from the social system. These dynamics are quite evident when considering on-line communities, such as those of e-Bay or Amazon's customers and sellers.

5.4b Information warfare

Information warfare is one of the most compelling examples of the effects of the information revolution on current society. The design of data banks and software, the ability to blindside an opponent's informational infrastructures, and ensure the superiority of informational infrastructures of a state, are as important as the superiority of weaponry and military force. This is the reason why, in the last two decades, several states have devoted huge effort and resources in order to improve their informational infrastructures and to educate experts in the relevant fields. ICTs prove to be effective and advantageous war technologies, as they are efficient and relatively cheap compared to the general costs of traditional warfare (Arquilla & Borer, 2007).

Information warfare is not only about using ICTs as new weapons: it is also about the need for states to establish their authority in the new domain, i.e. the non-physical one, described in section 5.3a. Information warfare engenders so many changes that it is deemed to reshape the very concept of war as we have known it for centuries. We may readily imagine that such changes also bring to the fore new and important ethical problems. In the rest of this section we will analyse the nature of information warfare and the ethical problems that it poses.

War is understood as the use of a state's violence through the state military forces to determine the conditions of governance over a determined territory (Gelven, 1994). The choice to undertake a (traditional) war usually involves a substantial commitment, as it has heavy economic and political costs, borne mainly by the civil society.

These features of war have been radically changed by information warfare, which provides the means to carry out war in a completely different manner. In this scenario, the changes brought about by information warfare are of astounding importance as they concern both the way militaries and politicians consider waging war, and the way war is perceived in civil society. Like traditional warfare, information warfare is very powerful and potentially highly disruptive; however, unlike traditional warfare, information warfare is potentially bloodless, cost-effective, and is not a military-specific phenomenon.

Let us first stress that it would be a mistake to consider information warfare simply as a non-sanguinary, cheap and less military-based version of traditional warfare. Information warfare can be as bloody and violent as traditional warfare, as it may determine damages and casualties comparable to traditional warfare. In other words, information warfare may range from a non-destructive phenomenon to a highly violent and bloody one. We may refer to this aspect as the *transversality* of information warfare.

Information warfare is transversal not only with respect to the level of violence, but also with respect to the domain in which it can be waged and the kinds of agents involved in it. Such transversality represents the ultimate difference between information warfare and traditional war, and it is the aspect of information warfare from which conceptual and ethical problems arise.

Let us consider *domain* transversality. In section 3.3a it was argued that, with the information revolution, the environment in which we act is extended to include both the physical and non-physical domains. Information warfare may originate in one domain and affect both of them.

The transversality of information warfare also has a bearing on the kind of agents involved in the warfare scenario. In this respect, two issues need to be highlighted: the ontological and the social status of the combatants. The ontological status ranges over quite a large spectrum, as combat actions undertaken in information warfare are performed both by artificial agents, such as viruses, drones and robots, and human agents. The heterogeneous nature of combat agents is an important aspect to consider when dealing with ethical issues. Typical problems concern, for example, ethical responsibility for the actions performed by artificial agents. In information warfare artificial agents and human agents may have the same role in achieving a given goal; their actions are equally relevant and important, despite their ontological differences. Therefore it is of paramount importance to define criteria for establishing the responsibilities in combat actions.

The transversality of information warfare with respect to the social status of the combatants follows from the fact that information warfare does not require military-specific skills and techniques. This aspect has the side effect of allowing skilful civilians to participate in combat actions in information warfare. Besides the image of a nerdy guy sitting in his room and blowing up a far distant nuclear power plant, this aspect of information warfare has an important consequence for current society, as it leads to the blurring of the distinction between civil society and military organisation.

When considered from an ethical standpoint this aspect of information warfare leads to new ethical issues such as whether it is acceptable from an ethical and political perspective to allow the distinction between military personnel and civilians to vanish, for this blurring of boundaries may eventually lead to holding civilians

responsible for combat actions and to considering civilian public and private infrastructure legitimate targets in warfare. So, for example, it may become acceptable to disrupt the civilian supply chain for food and water and to control civilians' private networks and computers.

Having analysed the nature of information warfare we can now provide a definition of this phenomenon.

> **Definition**. Information Warfare is the use of ICTs within an offensive or defensive military strategy endorsed by a state and aiming at the immediate disruption or control of the enemy's resources, and which is waged within the informational environment, with agents and targets ranging both on the physical and non-physical domains and whose level of violence may vary upon circumstances
> (Taddeo, 2011)

Now that we have clarified the nature of information warfare, we can move on and consider the applied ethical problems that it poses. Such problems are grouped under three categories on which both policy-makers and ethicists focus their attention. The three categories of problems are risks, rights and responsibilities. They can be referred to as the 3R problems. They are concisely described as follows.

> **Risks**. The risks involved with information warfare concern the potential increase in the number of conflicts and casualties. ICT-based conflicts are virtually bloodless for the army that deploys them. This advantage has the drawback of making war less problematic for the force that can implement these technologies, and therefore making it easier, not only for governments, to engage in ICT-based conflicts around the world so increasing the risk of escalation and therefore for casualties.
> (Arquilla & Borer, 2007)

> **Rights**. Information warfare is pervasive for it not only targets civilian infrastructures but may be launched through civilian computers and websites as well. This may initiate a policy of higher levels of control enforced by governments in order to detect and defend their citizens from possible hidden forms of attacks. In this circumstance, the ethical rights of individual liberty, privacy and anonymity may come under sharp, devaluating pressure.
> (Arquilla, 1999; Denning, 1999)

> **Responsibilities**. This category concerns the assessment of responsibilities when using semi-autonomous robotic weapons and cyber-attacks. In the case of robotic weapons, it is becoming increasingly unclear who, or what, is accountable and responsible for the actions performed by complex, hybrid, man-machine systems on the battlefield (Matthias, 2004; Sparrow, 2007). The assessment of responsibility becomes an even more pressing issue in the case of cyber-attacks, as it is potentially impossible to track back to the author of such attacks (Denning, 1999).

5.5 Conclusion

In this chapter we focused on the information revolution, its conceptual implications and the changes and the ethical problems that it engenders in current societies. We analysed two ethical problems brought to the fore by the information revolution: online trust and information warfare.

Online trust has been described as a property of the relations of communication occurring online. It has been stressed that it facilitates the occurrences of such relations and determines some advantages for the trustor, as it allows the trustor to achieve a goal while saving effort and commitment.

Information warfare has been defined as a new kind of war, which has the peculiarity of being transversal with respect to the domain in which it is waged, the nature and the social status of the agents deployed in it and the level of violence it may generate. The analysis concluded by describing three categories of ethical problems that information warfare poses and that are currently debated by ethicists and policy makers.

5.6 Exercises

1. Describe other ethical problems afflicting our society due to the information revolution.

2. Provide an argument in defence of the occurrence of trust in online interactions.

3. Indicate three more ethical problems that may concern the waging of information warfare and that have not been described in this chapter.

5.7 Suggestions for the exercises

1. Think about your studying and social activities of the past week and consider how many times you trusted an artificial agent and whether you trusted it more than a human agent.

2. Make a list of advantages and shortcomings that may follow from trusting artificial agents.

3. Compare your studying experience with that of a student from the previous generation. What are the aspects that changed due to the information revolution?

5.8 Further reading

For an overview of the most recent literature on online trust see Ess and Thorseth (2011). For a more in depth analysis of issues related to information warfare see Arquilla and Ronfeldt (1997). For a business-oriented analysis of ambient intelligence see
http://www.research.philips.com/technologies/projects/ami/breakthroughs.html#Context_awareness.

Part III: Knowledge and language

Chapter

6

6. MEANING

What does information mean?

6.1 Introduction

> *How can the semantic interpretation of a formal symbol system be made intrinsic to the system, rather than just parasitic on the meanings in our heads? How can the meanings of the meaningless symbol tokens, manipulated solely on the basis of their (arbitrary) shapes, be grounded in anything but other meaningless symbols? (Harnad, 1990, p. 335.)*

We mean many things by "information". The anthropologist and cyberneticist Gregory Bateson defined information as 'a difference which makes a difference.' (Bateson, 1972, pp. 320-321.) This might make information sound like almost anything. To some extent this is true. We are used to hearing the word "information" used in so many contexts from freedom of information, to the information age, to the information superhighway, to information overload and so on. This chapter looks more closely at the relation between information and meaning: both how information gets its meaning – in other words, what makes information meaningful – and how we can explain meaning using the conceptual tools of information theory. In previous chapters we have asked "What is information?" The question that concerns us in this chapter is how information can be meaningful.

6.2 The theory of meaning and the symbol grounding problem

The theory of meaning is an enormous topic in the philosophy of language that also reaches into many other areas of philosophy. In its most general form, it is the attempt to explain how our language connects to the external world; that is, the relation between what we say and what we are speaking about. The issue is not as straightforward as it sometimes appears at first glance. At the start of the twentieth century, Ludwig Wittgenstein wrote:

> The difficulty of my theory of logical portrayal was that of finding a connection between the signs on paper and a situation outside in the world. I always said that truth is a relation between the proposition and the situation, but could never pick out such a relation.
> (Wittgenstein, 1961, 19e-20e.)

We might intuitively think that the meaning of a symbol or word consists in the relation between that symbol or word and the thing to which the symbol or word refers. However, it's not at all clear what that relation is, especially when we consider how varied are the symbols that can be meaningful and how protean and fickle is the relation between symbols and meanings. Can an informational approach to meaning solve some of these age-old philosophical questions?

Consider, for example, the problem of how the same thing can have different names. The philosopher of language, Gottlob Frege, remarked that one of the planets (Venus) was known by two names: those who saw it at sunrise knew it as "the Morning Star," and those who saw it at sunset called it "the Evening Star." He concluded that the thing that a word refers to is not necessarily the same as the meaning of the word. Frege distinguished two aspects of meaning which he called *Sinn* (or "sense") and *Bedeutung* (usually translated as "reference") (Frege, 1892). Those two names – "the Morning Star," and "the Evening Star," – have the same reference – viz. the planet Venus – but a different sense. This reference-based theory of meaning was followed by numerous attempts to find the "hook" that connects our words to the things they are about. An informational approach widens this claim considerably. Theories that explain meaning have been limited by their focus on language, specifically on speech and writing. This informational approach tries to explain meaning by referring to a broader phenomenon than language: information. This gives rise to two central questions: (i) how information acquires its meaning (and consequently, to the question of what meaning is); and (ii) how the theory of meaning can be illuminated by concepts from information theory.

Let us take each of these in turn. The first question has been called the "meaning grounding problem" and it concerns where meaning comes from or how an entity can acquire meaning. We recognize some marks, symbols, sounds, gestures, signs, and so on as meaningful, yet others are meaningless. But just knowing the relationship between symbols or sounds does not ever amount to knowing the meaning of those symbols and sounds. Consider a subspecies of the meaning grounding problem which concerns how symbols or data acquire meaning: the symbol grounding problem. It can be illustrated with a familiar example. Suppose that you have just arrived in a foreign country with no knowledge of the local language and you are attempting to find your way to your hotel. You notice a sign with some letters printed on it, you hear people speaking and gesticulating, you see traffic directions and symbols. All of this is more or less unintelligible to you. You buy a dictionary but it is entirely written in the local language. Nevertheless, you persevere and look up the word on the sign. You then begin to look up the words in the definition; and then the words in that definition, and so on. Of course, you will never be able to derive any meaning from this dictionary as the mere relations between symbols cannot provide you with the "hook" that connects those symbols to the world which they seem to refer to.

Stevan Harnad used Searle's Chinese Room Argument (Searle, 1980) to demonstrate the symbol grounding problem (see also Chapter 11). The person in the Chinese Room appears to have no access to the meaning of the symbols it can successfully manipulate syntactically. Similarly, a robot or artificial agent (AA) will inevitably be unsuccessful since the mere physical shape and syntactic properties of a symbol provides no information as to its corresponding semantic value (Taddeo & Floridi, 2005).

Harnad described the problem thus:

> How can the semantic interpretation of a formal symbol system be made intrinsic to the system, rather than just parasitic on the meanings in our heads? How can the meanings of the meaningless symbol

> tokens, manipulated solely on the basis of their (arbitrary) shapes, be grounded in anything but other meaningless symbols?
> (Harnad, 1990)

For example, the symbols in the Venus case are the words themselves: "the," "Evening", "Star", "Morning" and the other English words written in the Roman alphabet that have meaning to a speaker of that language. The formal symbol system is a combination of a number of things but essentially includes these symbols and the rules required to form new complex symbols – combinations of the basic, atomic symbols e.g. 'the Evening Star is the same as the Morning Star'. Now suppose you are not an English-speaker and nor are you familiar with any Indo-European languages or alphabets. In that case, the symbols have no more meaning to you than the alphabets of any other language you do not understand. The problem is that there is nothing in the marks themselves or in their relation to other marks that can give you that meaning. If we do understand it, we are dependent on the meaning generated by our own minds. If we find meaning in the symbols – for example, the particular colour pattern before our eyes and the spatial relationships between the marks – then are we not just imposing the meaning that we already have in our minds onto the environment? If so, where did the meaning in our heads come from in the first place? We seem to run into a regress.

Some cognitive scientists and AI researchers have argued that cognition can be reduced to a kind of symbol manipulation. Consider the discussion of computationalism (Chapter 10.3). It discusses how a general model of cognition emerged out of von Neumann's three-stage computational model: *input→ processing→output*. Jerry Fodor called this the horizontal architecture of cognition (Fodor, 1983). In essence, the cognitive system receives an input from the environment; it then processes the input, and selects an action based on the input, current state, and its history. We can similarly describe this as a three-stage sequential pattern: *perceive→think→act*. Fodor argued that the meaning of symbols is grounded in the relationship between the formal symbol system and the world. If correct, this is promising as we can theoretically simulate cognition in artificial agents by equipping them with the right rules for manipulating symbols. In humans, this symbol manipulation is subconscious and autonomous and once the brain receives symbols, then, assuming it is functioning properly, it will acquire the relevant meanings. Recent research in cognitive science, however, highlights the interpretative and problematic nature of picking out the objects, events, and states of affairs that symbols refer to and so the 'symbolist' account seems to simplify the function of cognition. There is more to it than symbol manipulation.

So how ought we to resolve the symbol grounding problem? One option is to look more closely at the processes of perception. We can say that when we learn a language, for example, we learn how to appropriately map symbols to specific referents. Suppose that, in the illustration given above, you abandon your dictionary and try to work out what the people around you are saying. You listen to their words, observe what they point to, and construct a small bilingual dictionary of your own. Each time someone points to an object and appears to refer to it you write down what they say and translate it into your own language. Unfortunately, you sometimes hear someone say a word and seem to refer to a completely different object or situation. Again, no matter how close and detailed your observations, it seems that you will never completely understand what the meaning of the word is just by noting the structural relations between words, symbols and gestures.

What is more, whilst symbols are usually discrete, self-contained entities, the external world for the most part is not. If we go along with this approach we are either sorely lacking in symbols if we want to

describe the world well or we conflate many things when using the same symbols. What is more, as described by Hilary Putnam, just as a set of mathematical equations can be mapped onto a near-infinite number of structures in the world, a system of relations between symbols can be mapped onto a near-infinite number of structures in the world. Hence, there is no non-arbitrary way in which symbols acquire meaning (Lakoff, 1987; Putnam, 1981).

A second approach is to abandon our assumption that there is a direct relation between symbols and referents. Such an alternative approach might be found in research in embodied cognition, which takes cognition to be dependent on the physical capacities and actions of an agent. Specifically, we can view meaning as a way of coordinating action to achieve certain goals. According to this approach, the meaning of a particular situation for the agent is the combination of actions available to it. For example, the meaning of being close to a car, for a human, might be travel, but for a cat it might be shelter. Consequently, the symbol grounding problem is dissolved rather than solved. There is no need to look for the "hooks" that connect references and referents, speech and what we speak about. Rather, what ought to be matched is goals and affordances: what the agent wants to achieve and what opportunities are afforded by a situation to achieve that goal. This approach is considered further in Chapter 10 and we will return to it in this chapter in section 6.

6.3 Statistical approaches: Shannon and Weaver, Bar-Hillel and Carnap

Recall Shannon's landmark paper "The mathematical theory of communication" discussed in Chapter 1.3, and subsequent research by Shannon and Weaver. Shannon described an information-generation system with five essential components: an information source, a transmitter, a channel, a receiver, and a destination. (Shannon, 1948, p. 4; Wiener, 1948, p. 79). Their model defines what have become the main components of many subsequent mathematical or probabilistic approaches to information, such as those proposed by Bar-Hillel and Carnap, and Dretske discussed below. This model has been applied not just to describe the external processes between two cognitive agents but also the internal processes of the mind. Sperling (1963), for example, described how environmental sounds are processed through cognitive systems and transformed until they reach the conscious level. Shannon's analysis of information, however, is not an analysis of meaning but of the correlations between messages, variables, etc. (see Chapter 1.3). Although there have been further statistical models based on Shannon's work, let us look at the Bar-Hillel and Carnap model, which sought to do justice specifically to the problem of meaning which Shannon and Weaver had set aside.

Bar-Hillel and Rudolf Carnap developed a probabilistic approach and applied it to the meaning of propositions. Their approach was based on what is called the inverse relationship principle. According to this principle, the amount of information associated with a proposition is inversely proportional to the probability associated with that proposition. The core idea is that the semantic content of p is measured as the complement of the a priori probability of p,

$$\mathrm{CONT}(p) = 1 - \mathrm{P}(p)$$

where CONT is the semantic content of p. Crudely, CONT(p) is a measure of the probability of p *not* happening, or *not* being true. This means that the less probable or possible p is, the more semantic information p is assumed to be carrying. Tautologies, like "all ravens are ravens", have to be true. So they are assumed to carry no information at all. Since the probability that all ravens are ravens is 1, P(p) is 1, so CONT(p) is 1-1, i.e. 0. By extension, we might presume that contradictions – statements which describe impossible states or whose

probability is 0, such as "Alice is not Alice" – contain the highest amount of semantic information. Thus we seem to run into what has been called the Bar-Hillel-Carnap paradox: the less likely a statement is, the greater its informational content, until you reach a certain point at which, presumably, the statement contains no information at all since it is false. As Bar-Hillel and Carnap state:

> It might perhaps, at first, seem strange that a self-contradictory sentence, hence one which no ideal receiver would accept, is regarded as carrying with it the most inclusive information. It should, however, be emphasized that semantic information is here not meant as implying truth. A false sentence which happens to say much is thereby highly informative in our sense. Whether the information it carries is true or false, scientifically valuable or not, and so forth, does not concern us. A self-contradictory sentence asserts too much; it is too informative to be true
> (Bar-Hillel & Carnap, 1953, p. 229)

6.4 Probabilistic approaches: Dretske

Fred Dretske was one of the earliest philosophers formally to connect the Mathematical Theory of Communication (MTC) to the problem of knowledge. In a theory he called Indicator Semantics he wrestled with the problem of how data and information can be "upgraded" to knowledge. Dretske supposed that the relationship between data and meaning was central to resolving this puzzle and key to this relationship was the further connection between signal and receiver. As part of Dretske's effort to use the concept of information to explain and develop problems of knowledge, perception, and meaning, several other philosophically interesting issues are tackled. He firstly gives an account of the propositional content of a signal (events, structures, situations) and develops a semantic theory of information. For Dretske, information is an informee-independent, objective phenomenon that occurs in a multitude of ways and which existed before the development of agents (e.g. humans) with the ability to selectively utilise it. Regularity and *law-like* or non-accidental correlations are crucial to information flow; without a lawfully regular universe, no information would flow.

Dretske uses the MTC as a starting point before establishing its limitations and departing to develop his theory of semantic information for philosophical use. The first thing to note is that MTC is a syntactic treatment of information and is not, at least directly, concerned with the semantic aspects of information – with what the information *means*. It is really an account of data quantification and transmission, as opposed to a theory of information as it is commonly understood. The quantity of data a signal carries can provide *some* information about its content. For example, say that a jug will either contain 1 litre of milk or 2 litres of water. Learning that the jug contains only 1 litre of liquid provides information about what type of liquid the jug contains, namely milk. Ultimately, MTC is an insufficient tool for the analysis of semantic information. Secondly, MTC is concerned with the statistical properties of transmission, with the average amount of information generated by a source. However information as it is commonly understood, and for Dretske's purposes, is something associated with individual events. It is only particular signals that have content which can be propositionally expressed.

So Dretske therefore must adapt the ideas of MTC for his purposes. After some toil, he comes up with the following definition of a signal's (structure's) information content:[13]

[13] See Chapter 13.5 where Dretske makes an important distinction between 'information content' and 'semantic content'. See also Dretske (1981, pp. 171-189.)

A signal *r* carries the information that *s* is *F* = The conditional probability of *s*'s being *F*, given *r* (and *k*), is 1 (but, given *k* alone, less than *1*)

Here *k* is a variable that takes into account how what an agent already knows can determine the information carried (for that agent) by a signal. For example, Alice's utterance that she is from England does not carry the information that she is from London (for she could be from Birmingham, Manchester, etc.). But if you already know that Alice originates from a capital city, then for you her utterance does carry the information that she is from London. Although the *k* variable which occurs in Dretske's definition relativizes information to what the receiver already knows concerning the possibilities at the source, this relativization is only meant to accommodate the way information is thought about, that the information one can get from a signal depends on what one already knows; it does not undermine the essential objectivity of information.

As an example of how the use of probability works here take the information-bearing signals of a clock. The signal from a correctly functioning clock carries the information that the time is such-and-such. In accordance with the above definition, if one looks at a correctly functioning clock that reads 6:30pm, then the conditional probability that the time is 6:30pm given the clock's signal of 6:30pm is one. If the clock were malfunctioning, things would be different. Say that the clock stops working at 6:30pm and the next day someone happens to look at the clock when it is 6:30pm. Even though the time indicated by the clock happens to be the actual time, the clock here does not give the information that the time is 6:30pm because the conditional probability that it is 6:30pm given that the clock shows this is less than one.[14]

Here are three reasons given by Dretske for his stipulation that the value of the conditional probability in his definition of information be one and nothing less:

- One basic principle of information flow that should be satisfied is the conjunction principle: if a signal *A* carries the information that *B* and *A* carries the information that *C*, then it also carries the information that *B* and *C*. If the conditional probability requirement was relaxed and made lower than one, then this principle does not hold.[15]

- Another principle of information flow Dretske holds is what he terms the Xerox principle: if *A* carries the information that *B*, and *B* carries the information that *C*, then *A* carries the information that *C*. So information flow is *transitive*. Once again, if the conditional probability threshold is set to anything less than one, then this intuitive principle does not hold.[16]

- Finally, and as Dretske frankly acknowledges, there is no non-arbitrary figure at which to impose a threshold. If information can be obtained from a signal involving a conditional probability of less than one, then information loses its "cognitive punch". To use an example of Dretske's, think of a bag with ninety-four red balls and six white balls. If one is pulled out at random, you cannot *know* that it was red. Hence, why suppose you have the information that it is red?

[14] Since there are 1440 minutes in a day, the probability that the time is 6:30pm given the non-functioning clock's signal of 6:30pm is in fact 1/1440.

[15] Here is how this can happen: $\Pr(B|A)$ stands for the probability of *B* given *A*. Where *x* stands for a probability value below 1 that is the minimum threshold at which the information relation would hold, according to probability it can be the case that $\Pr(B|A) \geq x$ and $\Pr(C|A) \geq x$ but $\Pr(B \,\&\, C|A) < x$.

[16] Once again, where *x* stands for a probability value below 1 that is the minimum threshold at which the information relation would hold, according to probability it can be the case that $\Pr(B|A) \geq x$ and $\Pr(C|B) \geq x$ but $\Pr(C|A) < x$.

A major concern against setting the required conditional probability to one is that, since there are very few conditional probabilities of one out there, very little information ever flows. Dretske deals with these concerns by introducing the idea of *fixed channel conditions* and *relevant alternatives*. The basic idea is that the conditional probability requirements are made relative to a set of possibilities relevant to the communication channel. This is a more flexible and realistic way of thinking about the conditional probabilities defining information flow. That calculations are made within a background of stable or fixed circumstances is not to say that circumstances cannot change. Rather, it is only to say that for the purposes of determining conditional probabilities, if conditions are normal and there is no significant chance of something happening, such changes are set aside as irrelevant.

For example, take our previous scenario in which a functioning clock is accurately correlated with the time and is thus transmitting information about the time. Now, there are scenarios alternative to this one, such as cases where the clock has flat batteries, in which the clock's signal is not correlated with the time and it therefore fails to transmit information about the time. If these alternatives are factored into the probability calculations, then the probability that it is time x given that the clock shows x is not equal to one. Consequently, in the actual scenario the clock would not meet the requirements of Dretske's information flow definition and so would be judged as not carrying the information that the time is such-and-such. But the idea is that, if the batteries in the clock are in good working order and the clock is functioning correctly, even if there is a minute probability that the new batteries could become defective or a mechanism in the clock could break just before someone looks at the clock, these possibilities are ignored in calculating the information the clock is delivering. Possible (non-actual) but far-flung or improbable alternatives such as these are deemed irrelevant in considering the set of relevant alternatives against which probability calculations are made. Even if there is technically a non-zero probability that they could occur, if these issues have not actually occurred, then they are dismissed.

There is no determinate method to decide what counts and what does not count as a relevant alternative. In general the selection will depend upon the knowing agent and their environment and will also be a pragmatic decision. Particularly due to this lack of clear determination the notion of relevant alternatives has been a point of philosophical contention and from one point of view its application is seen to be somewhat *ad hoc*. Nonetheless, it is a valuable idea that can serve as a foundation for accounts that afford a way to realistically talk about information (and knowledge).

One piece of support for this strategy of employing channel conditions/relevant alternatives can be found in analysing the application of absolute concepts. The concept of information here, like the concept of knowledge to which it will be applied, is absolute: A either carries the information that B or it doesn't. Now, we legitimately apply absolute concepts all the time, even though at some extremely magnified level they might fail. For example, we might say that an apple box is empty because it contains no apples, even though in some sense it is not empty since it still contains dust and air molecules. Of course, when attributing emptiness to the box, we rightly do not include dust and air molecules in our consideration; they are "irrelevant alternatives" as it were. In this way, we can legitimately apply absolute concepts.

It is important to note that a signal's informational content is not unique. Generally speaking, there is no single piece of information in a signal or structure. For example, anything that carries the information that x is the number 7 also carries the information that it is an odd number, that it is a prime number, that it is not an even number and so on. This non-uniqueness is one characteristic of information that distinguishes it from meaning; if I say that x is the number 7, I mean just that. Also, whilst the meaning of

a signal is independent of its truth, its information is not. Thus information and meaning are two distinct things and it is important that the two are not confused. In Chapter 3 we also examined how Dretske's theory relates to his theory of how symbols acquire meaning based on information-carrying tokens.

6.5 Semantic approaches: Levels of abstraction, and syntax

The fourth approach we will look at is known as the General Definition of Information (GDI) (Floridi, 2005b). This theory proposes a tripartite definition of information:-

The General Definition of Information (GDI)
p is an instance of information, understood as semantic content, if and only if:
(GDI.1) *p* consists of one or more *data*;
(GDI.2) the data in *p* are *well-formed*;
(GDI.3) the well-formed data in *p* are *meaningful*.

In other words, if what you have is not a well-formed combination of meaningful data, then you do not have information.[17] Here, "well-formed" means that the data are correctly structured according to the relevant syntax. The question that falls out of this statement – viz. when is data correctly structured? – is problematic enough. We have already discussed the importance of a relevant syntax. The question raised by (GDI.3) is the primary concern of this chapter – namely, when is well-formed data meaningful? In other words, how can data acquire their meaning? No doubt, we often attribute meanings to data. We can point to a datum on a graph and infer that the datum means that, say, the recorded temperature at a certain pressure is 50°C. We can read a weather report and infer that the data in the report mean that it is likely to rain tomorrow. But what are needed are general principles which would be exportable across conceptual platforms. What allows us to make these inferences is the structure of the data and the syntax that we use to interpret it. Here, syntax means something broad. It is the system, structure, code or language that determines the form, construction or composition of information. It is not necessarily linguistic – think, for example, of an instruction diagram for building flat-pack furniture, or Morse code signals. Bateson, in the work quoted at the head of this chapter, notes that on a map, only differences in altitude, and not consistencies in altitude, would be represented (i.e. as contours). But if we do not understand the syntax of the contour lines – the differences – we may misunderstand what they represent.

Consider the following illustration from Dretske: an airplane pilot uses altimeters to determine the plane's altitude. An altimeter is essentially a pressure gauge – it responds to changes in air pressure and the pilot takes this to represent changes in altitude. Note that we must say "takes this to represent" here as what is crucial is not just what the device measures but what those measurements are taken to represent. Conflating the two can lead to bad inferences being drawn from the altimeter's movements. For example, if the aeroplane strays into an environment that has an unusual air pressure, and the pilot takes its movements to represent altitude, he will acquire false beliefs about the altitude. Similarly, if we take the altimeter out of the plane and place it in a depressurized container, the altimeter will register a much higher altitude than is, in fact, the case. In such examples, it is not that the altimeter is malfunctioning – it is doing its job perfectly well – but that we have incorrectly understood the syntax of the information it provides. We do not understand what this information means. When the altimeter responds to external

[17] Note that this is not Floridi's view due to his veridicality thesis (see Chapter 7.2). Also see Floridi (2010c, pp. 227-258) for more on Floridi's approach.

changes in its environment it is not actually generating any information *per se*. What is generated might instead be called data. We need a syntactical structure – what philosophy of information calls a Level of Abstraction (see Chapter 2) – to give that data meaning and determine its informational content. We often hear people say, and have no doubt said ourselves, things like, "It looks cloudy. Does that mean it will rain?", "What does irascible mean?", "That's the third time a black cat has crossed my path. What does it mean?" and so on. What we are saying when we use the word "mean" is given by the kind of answers that we expect to satisfy the question. The answer that clouds are often used in literature to symbolise the coming of rain would not be a helpful answer to the first question, but that the presence of certain types of cloud indicates that it is likely to rain would. The answer that it means that a feline animal with dark fur has travelled in such a manner as to move adjacently to your own trajectory would not be a helpful answer to the last question, whereas the answer that black cats crossing your path give you luck might be satisfying (to the superstitious amongst us). We might conclude from this that the meaning of a bit of information depends on our interests or goals.

6.6 Pragmatic approaches

We receive a lot of stimuli or signals in our day-to-day lives – either through newspapers, television, and the internet, or just the stimuli and signals we receive from all the objects we pass as we walk down the road. Our minds need a way of managing all that data and selecting only the meaningful, informative, and – hopefully -- true. One of the first steps in this process is deciphering what this putative information means. Charles Sanders Peirce was an American philosopher in the late nineteenth century who co-founded the Pragmatist movement with William James and John Dewey, and who also provides insight into the problems we face in deciphering the meaning of various signals (De Tienne, 2006). Peirce thought that this problem of the meaning of signals, what he called *Semiotics*, was in fact something of a *prima philosophia*. He wrote, '[I]t has never been in my power to study anything, – mathematics, ethics, metaphysics, gravitation, thermodynamics, optics, chemistry, comparative anatomy, astronomy, psychology, phonetics, economics, the history of science, whist, men and women, wine, metrology, except as a study of semiotic.' (Peirce, 1977, pp. 85-86). Clearly, Peirce thought semiotics could help with quite a range of studies!

Recapitulating Peirce's work on information into a complete account is something that has to be done through understanding a series of lectures, manuscripts, texts, and notebooks as Peirce himself did not compose a summation of his work in any one text. For our purposes, we will focus on the key elements of his theory of information and how it relates to the theory of meaning. Central to this is his analysis of propositions which he argued had two qualities: extension and comprehension. Peirce argued that any proposition consisted of an ordered triplet of references:

1. A direct reference to its object (the real things that it represents),

2. An indirect reference to the characters common to these real things, and

3. An indirect reference to an interpretant defined as the totality of facts known about its object.
(De Tienne, 2006)

For Peirce, each of these references is "informational" in nature. The first refers to the "informed breadth" of the proposition, the second to the "informed depth" of the proposition, and the third to the "information" concerning the proposition. By "informed breadth", Peirce meant that the proposition must be predicable of real things, 'with logical truth on the whole in a supposed state of information.' (Peirce, 1984, p. 79.) "Informed

depth" is measured not by the number of "mere names" that can be attached to the subject, but to the number of distinct properties that can be said to belong to the subject by induction. Information comes about as the product of these two logical quantities, breadth and depth.

Peirce argued that in order to understand the meaning of a signal, we are not concerned with every aspect of the signal but only the *signifying element*. For example, in order to understand that smoke means fire, we do not need to know everything about the smoke – its shape, the particular way its fumes form, its precise colour, and so on – but only that element of it that signifies the presence of fire. Peirce used a lot of different terms for the signifying element of a signal – "sign", "representamen", "representation", and "ground". We will just call them signs. Consider a beehive in your garden as a sign that there are bees in your garden. It is not every single characteristic of that beehive that signifies that there are bees in your garden. The colour, size, or shape of the hive is not particularly important and plays what Peirce calls a "secondary signifying role". The primary signifying role in this case is the causal connection between the type of object that a beehive is and the presence of bees. This relationship is the sign. The meaning of this sign is "there are bees here"! There may be other signs that there are bees in the garden: the pollen count, a bee-sting on a child that unfortunately bothered one of the garden's residents, and the noise that bees make perhaps. What makes these things signs is their capacity to indicate the presence of bees. The colour of the stripes on a bee, its gender or age, are not essential to indicate the presence of the bee and so are not signs that there are bees. A second element to this connection is, in Dretske's terms, that receiver's interpretation of the sign. Roughly, this is the meaning we take from the relationship between the sign (the beehive) and the object (the bees). Peirce thought that signs determine their interpretation. That is, the beehive draws our attention to the connection between beehives and bees and in so doing determines that we will believe that there are bees.

In his later writings, Peirce considered the conditions that synthetic propositions must meet in order to be considered genuinely informative. (See De Tienne (2006) and Robin (1967, pp. 9-10).) He proposed the following five conditions:

1. An informative proposition must convey a truth. That is, it must not be dependent on human ideas and thoughts but on a real event independent of our interpretation.

2. The truth conveyed must not be novel. That is, the listener or receiver of the information must be affected by it in some non-trivial way.

3. The truth conveyed must be relevant. It must relate to a universe that actually concerns the listener or receiver.

4. The truth conveyed must not only be of interest to the receiver but must actually drive it to generate further interpretants sharing a similar purpose.

5. Finally, the truth conveyed must involve an actual possibility. The information is being stated not just for the sake of it but for the sake of the consequences that are entailed by it. Stating, for example, that 'This bridge is weak' is informative because it entails that it will break if sufficient pressure is exerted.

A more modern pragmatic theory of information can be found in MacKay (1969). MacKay defined information as 'that which does logical work on the organism's orientation' (MacKay, 1969, pp. 95-96). Under this view, information is related to an organism's cognitive structures as it reacts to its environment. It is consequently a much broader approach to information than, for example, Shannon information, since it also

includes natural sources of information. Some of these more recent pragmatic theories are discussed in Chapter 10.5a.

6.7 Intention-based semantics and computational systems

In section 6.6, we considered whether the meaning of information depends, in some sense, on what kind of answer we might expect to satisfy a question. Call this a causal account of meaning. Formally, a causal account states that "x is meaningful iff x tends to produce such and such a psychological state in a hearer and to be produced by a corresponding state in a speaker." The philosopher of language H. P. Grice argued against such an account. He noted that, for example, my picking up an umbrella tends to make whoever sees me think that it is raining outside and I may pick up an umbrella because I think it is raining outside. However, my picking up an umbrella does not *mean* that it is raining outside. Grice argues that the causal account ignores the "intentionality" of meaning. That is, the important element is not the effect that tends to be produced but the effect that the speaker intends to produce. Secondly, the causal account relies on a conventional or common meaning always being the meaning, whereas in reality speakers may mean many things by the same utterance on various occasions. Instead, Grice subscribed to intention-based semantics (Grice, 1957). According to this theory, the meaning of a sentence is determined by the psychological state it is *intended* to produce in the hearer. He distinguished between natural (non-cognitive) and non-natural (communicative) meaning. Natural meanings are those that, roughly speaking, could be given a naturalistic explication such as "That body temperature means that she has a fever," or "The code doesn't mean anything to me, but to the computer programmer it means that there is a software error." Non-natural meanings include those that cannot be given naturalistic explications such as "Those three rings on the bell mean that the bus is full," or "When Bob said that he didn't have a leg to stand on he meant that he had no support for his position."

Grice proposed two tests for each kind of meaning: an entailment test and a quotation test:

> **Entailment.** In cases of natural meaning, x *means that* p entails that p. In cases of non-natural meaning, x *means that* p does not necessarily entail that p. For example, "That body temperature means that she has a fever," entails that she has a fever. "That body temperature means that she has a fever, but she hasn't got a fever" is self-contradictory.
>
> **Quotation.** In cases of natural meaning, the verb "mean" cannot be followed by a quotation. In cases of non-natural meaning, it can. For example, one cannot write, "The code doesn't mean anything to me, but to the programmer it means that 'there is a software error.'" To write this might be to imply that you were being sarcastic and there really was no software error. On the other hand, one can write, "When Bob said that he didn't have a leg to stand on he meant that "he had no support for his position.""

When we talk about meaning in this communicational form, we often imply two concepts: intention and understanding. When Alice tells Bob that she is currently in Edinburgh, Scotland, she intends to transmit the information contained in what she says. She also intends that Bob is her audience and she assumes or predicts that he will understand what she says. Furthermore, in order for Bob to receive this information he must understand what Alice meant, at least to some degree. If he understands by her sentence something radically different from what Alice intended for him to understand then there has been a "failure to communicate": a failure, in other words, to transmit meaningful information.

In Chapter 8 we discuss themes in A.I. that draw parallels between computers and our own brains and minds. We can consider that question as it relates to meaning and the capacity or lack of capacity to mean what one

says: can computers mean what they say in the same way that humans do? Does meaning require intention? If computers can mean what they say or display then we might be more tempted to say that they are functioning like human brains (and vice versa if they cannot). If you find Grice's account convincing, you may want to ask whether a computer can intend to say things. And if you prefer the causal account you may want to think about whether the counterexamples apply to computers that purport to transmit meaning.

6.9 Conclusions

In this chapter we have looked at research which tries to answer the question of how data or information acquires meaning. We looked at statistical approaches such as Shannon and Weaver's mathematical theory of communication; probabilistic approaches such as Bar-Hillel and Carnap's or Dretske's theory of meaning; semantic and pragmatic approaches. We also looked at several specific problems in the philosophy of information such as the symbol grounding problem and the Bar-Hillel-Carnap paradox. Finally, we looked at intention-based semantics and asked what it would mean for non-animal agents to communicate or transmit meaning.

6.10 Exercises

1. What is the difference between the question of whether information is meaningful and the meaning grounding problem? Describe in your own words the symbol grounding problem.

2. What is paradoxical about the inverse relationship principle?

3. What kinds of things can carry an informational signal?

4. Think of three advantages that pragmatic approaches have over probabilistic approaches. What disadvantages are there to the pragmatic approach?

6.11 Suggestions for the exercises

1. Consider the following as two separate questions: is information meaningful? What is the meaning grounding problem? Now consider how the two differ in terms of the problems they aim to solve. Re-read section 6.2 to remind yourself of the symbol grounding problem. Be careful about the subtle difference between the meaning grounding problem and the symbol grounding problem.

2. Re-read 6.3. Recall the definition of the inverse relationship principle and consider why it may be called a paradox. Is it really paradoxical?

3. Write a list of things that can carry an informational signal, or, in other words, provide information to you or others. These are sometimes called 'information vehicles'. What qualities do these things share that might group them together as information vehicles?

4. Familiarize yourself with the probabilistic approaches and the pragmatic approaches. You may need to use the further reading to tackle this problem fully. Think about what problems in the philosophy of information each tries to solve and how successful they are in doing so.

6.12 Further reading

Dretske (1981), Floridi (2005b), MacKay (1969), Peirce (1977), Wilson and Foglia (2011).

Chapter

7

7. TRUTH

Must information be true?

7.1 Introduction

Suppose Alice asks Bob, "What is the capital city of France?" Bob, who is prone to lying, responds with "Rome." Has Bob provided Alice with a genuine piece of information?

In Chapter 6 a general definition of semantic information (GDI) as well-formed, meaningful data (semantic content) was given. Another aspect of semantic information to consider is the *alethic* (of or relating to truth) nature of information: does semantic content need to be true in order to qualify as semantic information, or does any semantic content, true or false, count as information? According to Fox: "*x* informs *y* that *p*' does not entail that *p* [and since] ... we may be expected to be justified in extending many of our conclusions about 'inform' to conclusions about 'information' [it follows that] ... informing does not require truth, and information need not be true.' (Fox, 1983, pp. 160-161, 189, 193.)

Fox thus advocates some form of the Alethic Neutrality (AN) principle: meaningful and well-formed data qualify as information, no matter whether they represent or convey a truth or a falsehood or have no alethic value at all.

In the discussion that follows we will not consider factual semantic content that has no alethic value (has no truth value, so is neither true nor false). The prime issue here is whether or not information requires truth.[18]

7.2 The veridicality thesis

According to AN, since information does not have to be true, semantic content already qualifies as semantic information. So the GDI is sufficient for information – information is well-formed, meaningful data. However, there has recently been some debate on the alethic nature of semantic information that questions whether these conditions are sufficient. This debate was initiated by Floridi's advocacy of a

[18] Although it is not the focus of this chapter, the nature of truth itself is a central subject in philosophy. Put simply, the problem of truth concerns determining what truths are and what, if anything, makes them true (Glanzberg, 2013). Correspondence, coherence, and pragmatist theories are the main theories of truth. In 'Open Problems in the Philosophy of Information' (Floridi, 2004b), one of the open problems is 'Can information explain truth?' Floridi (2010e) develops an alternative theory of truth for semantic information, namely a 'correctness theory of truth'.

veridicality requirement for semantic information – he claims that information must be true.[19] According to the veridicality thesis (VT), in order for semantic content to be counted as information it must also be true: semantic information is well-formed, meaningful and *veridical (truthful)* data. In other words, only true propositions count as genuine semantic information; false semantic content or "misinformation" is not genuine information. Bear in mind that this veridicality requirement applies only to factual semantic content, such as "that is a pizza" and not *instructional* semantic content, such as "go and make a pizza" which is neither true nor false – it is not *alethically qualifiable*.

Other notable advocates of a veridicality condition for information are Dretske (1981), Barwise and Seligman (1997), Graham (1999), and finally Grice (1989), who offers the following direct characterisation of this position: 'false information [misinformation] is not an inferior kind of information; it just is not information' (Grice, 1989, p. 371). Thus, according to VT, the prefix "mis" in "misinformation" is treated as a negation – "misinformation" is "not information".

There may very well be no objective fact about the world that could decide whether the veridicality thesis is true. But whilst it might seem that the debate is just a trivial terminological one, there is arguably more to it. Each side has reasons to offer and one's position can be motivated by what one wishes to do with the resulting notion of information.

Those who reject VT might say that, since semantic content in general plays a role in the cognitive activities of semantic agents, it should be classed as information. Suppose that Bob says to Alice that Rome is the capital city of France. Alice can process this false semantic content and use it. She can use it to answer (incorrectly) a quiz question asking for the capital city of France. Or she could use it to deduce unsoundly the information that Rome is a European capital city.[20] Denying that false information (misinformation) is a genuine type of information would unnecessarily restrict the range of cases in which the term "information" is legitimately applied.

Those who support VT might say that they want an account more in line with a certain ordinary conception of factual information, the sense in which information is a success word.[21] Data is data, semantic content is semantic content, false semantic content takes the term "misinformation" and true semantic content takes the term "information".

There are three positions one could adopt in this debate:

A. Genuine semantic information requires truth;
B. Any legitimate conception of semantic information will not have truth as a requirement; or
C. There is more than one legitimate conception of information. Some require truth and others don't.

By the end of this chapter you will hopefully have a better idea of the issues involved and which position you prefer.[22]

[19] See Floridi (2010c, Chapter 4) for more. Floridi himself acknowledges that this veridicality requirement is hardly a novel idea (some precedents are listed below). Nonetheless, (re)ignition of the debate shows that discussion of the issue remains to be had.
[20] Although the conclusion that Rome is a European capital city is true, the deduction is unsound because one of its premises was false.
[21] A success word is a word whose application to an embedded proposition implies the truth of that proposition. For example, I remember, know, realize, perceive that *p* all imply the truth of *p* (Blackburn 1994).
[22] Chapter 8, on information and knowledge, will also help to elucidate this matter.

7.3 Arguments for the veridicality thesis

Since the VT camp adds the extra condition that needs to be argued for, we will proceed by outlining some of their reasons intermixed with reasons and responses from the non-VT camp.

VT advocates maintain at least several technical or practical reasons for adopting VT. One prime reason is that a veridical conception of semantic information will provide an informational basis for *knowledge* (see Chapter 8). A core condition in the definition of propositional knowledge is that if one *knows* a proposition *p*, then *p* is true; one cannot know what is false. Alice can know that Paris is the capital of France but she cannot know that Rome is the capital of France. Since knowledge requires truth, attempts to define knowledge in terms of information will benefit from a conception of information that also requires truth.

Fred Dretske's work exemplifies this rationale. Under his information-theoretic epistemology, information is a necessary element of knowledge, so his aim is to formulate a theory of information where information, like knowledge, entails truth. He intends to respect some ordinary intuitions about what information is, where meaning (semantic content), which need not be true, is distinguished from information, which must be true. In his own words:

> As the name suggests, information booths are supposed to dispense information. The ones in airports and train stations are supposed to provide answers to questions about when planes and trains arrive and depart. But not just any answers. True answers. They are not there to entertain patrons with meaningful sentences on the general topic of trains, planes, and time. Meaning is fine. You can't have truth without it. False statements, though, are as meaningful as true statements. They are not, however, what information booths have the function of providing. Their purpose is to dispense truths, and that is because information, unlike meaning, has to be true. If nothing you are told about the trains is true, you haven't been given information about the trains. At best, you have been given misinformation, and misinformation is not a kind of information any more than decoy ducks are a kind of duck. If nothing you are told is true, you may leave an information booth with a lot of false beliefs, but you won't leave with knowledge. You won't leave with knowledge because you haven't been given what you need to know: information.
> (Dretske, 2008, p. 2)

As pointed out by Dretske, other disciplines such as the computing and information sciences freely employ the term "information" to refer to data or statements in general. In these cases truth seems to be irrelevant and anything that can be processed or stored in a database is counted as information. One reason is:

> to maintain that nonnatural [semantic] false information is information too mirrors our reason to posit nonnatural information in the first place: it allows us to capture important uses of the term "information". It is only by tracking such disparate uses that we can make sense of the central role information plays in the descriptive and explanatory activities of cognitive scientists and computer scientists, which partially overlap with the descriptive and explanatory activities of ordinary folk. (Scarantino & Piccinini, 2010, p. 323)

While this may be the case, bear in mind that "important uses of a term" are not necessarily justified or correct uses of the term. The term "vegetable" has important culinary and cultural uses, although

technically it is incorrectly applied in some cases. For example, botanically speaking aubergines (eggplants) and tomatoes are fruits, not vegetables. So perhaps it is preferable to develop a concept and terminology hierarchy in which data, semantic content and information are distinct. Under such a framework, cognitive, computer and information scientists would, generally speaking, traffic in data and semantic content. Although for computational purposes the data "Germany is in Europe" and "Mexico is in Europe" might be indistinguishable (they will be input, stored, manipulated and retrieved in the same way), it does not follow from this that the true datum counts as information if and only if the other does. Such an approach to information:

> blithely skates over absolutely fundamental distinctions between truth and falsity, between meaning and information. Perhaps, for some purposes, these distinctions can be ignored. Perhaps, for some purposes, they should be ignored. You cannot, however, build a science of knowledge, a cognitive science, and ignore them. For knowledge is knowledge of the truth. That is why, no matter how fervently you might believe it, you cannot know that Paris is the capital of Italy, that pigs can fly or that there is a Santa Claus. You can, to be sure, put these "facts", these false sentences, into a computer's database (or a person's head for that matter), but that doesn't make them true. It doesn't make them information. It just makes them sentences that, given the machine's limitations (or the person's ignorance), the machine (or person) treats as information.
> (Dretske, 2008, p. 2)

Michael Dunn responds to the line of argument exemplified in the above quotes from Dretske with a clever counterexample:

> I have heard a similar defence in a story of the "Information Booth" in a railway station and how it would be misnamed if it gave out false information. But note that I said "false information" in a very natural way. I think it is part of the pragmatics of the word "information" that when one asks for information, one expects to get true information, but it is not part of the semantics, the literal meaning of the term. If there is a booth in the train station advertising "food", one expects to get edible, safe food, not rotten or poisoned food. But rotten food is still food.
> (Dunn, 2008, p. 582)

As noted by Dunn, "false information" can be said in a very natural way. So what should we make of the term "false information"? Well, Dunn's point can be addressed with an argument made by Floridi (2005b), which offers a way to explain how it is that "false information" can be said in a very natural way whilst adopting VT and maintaining that "false information" is pseudo-information. The crucial distinction to make is that between *attributive* and *predicative* uses of "false". Whilst there are some technicalities to this strategy, its gist is easy to appreciate. Consider the following two compound terms:

1. false proposition
2. false economy

A false proposition is still a proposition; it is something that is both false and a proposition. On the other hand, when we say that an action, say, buying cheap shoes which soon fall to pieces, is a false economy, we are in effect saying that ultimately it is *not* an economy. Similarly, a false start in a race is not really a start.

With 1 "false" is being used predicatively to describe a property of the proposition. With 2 "false" is being used attributively to negate its subject; in effect it means "not an economy". Given this distinction, the idea is that we treat the term "false information" like "false economy": when we say that something is false information, we are in effect saying that it is not information.

It might strike the reader that this argument does not really settle much. For example, take the proposition "the earth has two moons". According to Floridi's analysis, this is a false proposition in the predicative sense and a piece of false information in the attributive sense. But this

> requires the brute intuition that that the earth has two moons is not information. The content of this intuition is nothing but an instance of the general thesis to be established. Thus, the argument is question-begging. No independent reason to reject instances of false information as information is given. Whether false information passes [this test] depends on whether we accept that a false p can constitute [semantic] information. We do!
> (Scarantino & Piccinini, 2010, p. 321)

This seems a fair point. Furthermore, note that there is a difference between "false economy" and "false information". With "false economy", we have established that "false" is used attributively and is in effect a negation. If "false" were to be used predicatively, then this would be a category mistake, as "false" is being applied to economies, which cannot really be true or false, but are instead real economies or not economies at all. However if false were to be used predicatively in "false information", this would not necessarily be a category mistake, since information as semantic content is the kind of thing that can be true or false.

So the claim that "false" is attributive as opposed to predicative when it is applied to "information" is unsettled and does not provide a conclusive argument for VT in its own right. It does though provide a way for proponents of VT to legitimately account for the term "false information". Given such a definition adhering to VT, "information" would be more like "tautology". In logic a tautology is a type of statement that is always true. For example, statements of the form "A or not-A" are instances of tautologies. Replace A with any proposition, for example "Bob is in Australia", and you will come out with a statement that is always true; "Bob is in Australia or Bob is not in Australia". "False tautology" in the predicative sense is a contradiction in terms, since tautologies are by definition true. With "false tautology" in the attributive sense, this means "not actually a tautology", in the way that an intuitionist might say that *A* or *not-A* is a false tautology.[23]

As we have briefly mentioned (and as the following chapter will discuss in detail), one reason to embed truth into the definition of information is that it makes it easier to give a definition of knowledge in terms of information. On top of this connection between the truth of information and knowledge, some definitions of information and being informed that are used for knowledge imply truth further, due to another requirement. According to these definitions, holding true semantic content is insufficient for being informed. To hold some piece of information *p*, it is important that the true semantic content "that *p"* was acquired in a reliable way that in some sense guarantees the truth of *p*. For example, imagine that Alice guesses the number of jellybeans in a jar to be 214 and luckily turns out to be correct. Although Alice's belief that there are 214 jellybeans is true, since it was not generated via a reliable method and could easily have been false, the content of her belief is not information and she has not been informed.

[23] Intuitionistic logic (Moschovakis, 2010) denies the law of excluded middle, the law which says that A v ~A (A or not-A) is always true.

On the other hand, if she emptied the jar and counted each jellybean, then since this method guarantees that she comes to the correct result, the true belief's content counts as information and she is informed that there are 214 jelly beans in the jar. Another way to put this is that to be informed that *q*, one must receive some signal which carries the information that *q*. If *p* carries the information that *q*, then every time *p* happens *q* happens; the occurrence of *p* guarantees the occurrence of *q*. As we shall see in Chapter 8, one form of information-theoretic epistemology incorporates such a definition of information and being informed in order to secure its definition of knowledge.

This relates to another point that the VT advocate could make. Floridi stresses the "relational" nature of information by comparison with food. Something does not count as food in general, but only *for* a certain type of organism. In this sense we could claim that food that isn't edible by some organism just isn't food for that organism. Similarly, misinformation just isn't information for a cognitive agent interested in knowledge.

7.4 Reasons to reject the veridicality thesis

Beyond using information to define knowledge, there are other reasons to consider VT that have to do with the satisfaction of certain intuitions and adequacy criteria one might expect from a definition of information. We shall look at some of these soon. But firstly let us look at some active reasons for not accepting VT.

Recall the Alethic Neutrality (AN) principle: meaningful and well-formed data qualify as information, no matter whether they represent or convey a truth or a falsehood or have no alethic value at all.

From an unrestricted Alethic Neutrality principle it follows that:

TA) tautologies qualify as semantic information
FI) false information or misinformation (including contradictions) are genuine types of semantic information, not pseudo-information

The acceptance or rejection of TA could go either way. Recall that tautologies have to be true. So tautologies are not informative in that they do not provide any new information about the world; but neither do they misinform. Furthermore, in some sense tautological deductive inferences can also be said to yield information. So whilst one option is to reject TA because tautologies are never informative, it is perhaps also reasonable to represent tautologies as 'instances of information devoid of any informativeness'. (Floridi, 2007a, p. 36).

Apart from passively defending FI based on it being the side that does not require a modification to the general definition of information as semantic content, there are some specific reasons for supporting FI. The following sample list is from (Floridi, 2005b):

False information can include genuine information
False information can entail genuine information
False information can support decision-making processes
If false information does not count as information, what is it? Assuming that p is false 'if S only thinks he or she has information that p, then what does S really have? Another cognitive category beyond information or knowledge would be necessary to answer this question. But another cognitive category

> is not required because we already have language that covers the situation: S only thinks he or she has knowledge that p, and actually has only information that p.'

In advocating VT, Floridi (2005b) argues against these reasons for supporting FI.

Treating false semantic content as information does lead to some counterintuitive results and certain problems. One issue concerns the quantification of information. With one standard account, semantic information is treated as semantic content (true or false) and the quantity of information given by a statement is inversely related to its probability; the less probable a statement, the more informative it is. In most cases this is plausible. Given the roll of a standard six-sided die, the statement "the die landed on 4" is more informative than "the die landed on an even number" since the former is less probable than the latter. Taking this to one extreme, since the tautological statement "either the die landed on 4 or it did not land on 4" has a probability of one, it has an information measure of zero. However taking this to the other extreme, since contradictions have a probability of zero, they are problematically assigned maximal informativeness. It seems rather bizarre to say that the statement "Paris is the capital of France and Paris is not the capital of France" yields more information than the statement "Paris is the capital of France and Berlin is the capital of Germany". VT paves the way for a method to measure information quantitatively in terms of verisimilitude and thus avoid such issues (D'Alfonso, 2011; Floridi, 2004c).

Contradictions, such as "Paris is not Paris" are necessarily false. They are a general problem for FI. One option is to exclude them, to say that only contingently false statements count as information. This strategy however does not completely free FI of contradiction-related issues.

An important question is how to measure the informativeness of contradictions. One standard practice is to assign them an informativeness measure of zero. But take the following principle of information aggregation, according to which the informativeness [info()] of two combined pieces of information is never lower than the informativeness of either single piece: if I_1 and I_2 are instances of information, then $\text{info}(I_1 + I_2) \geq \text{info}(I_1)$ and $\text{info}(I_1 + I_2) \geq \text{info}(I_2)$. Now, if I_2 is the negation of I_1, then their addition forms a contradiction. But if contradictions are assigned a measure of zero, then this principle of information aggregation is violated. Thus, in order not to violate this principle of information aggregation, it seems that the info() measure would have to be a partial function such that info(*C*) is undefined when *C* is a contradiction.

On a similar note, it is fair to say that information follows a principle of conjunction: for any two propositions *A* and *B*, if *A* is information and *B* is information, then the compound proposition *A & B* is information. However, this principle, together with the modified FI leads to the problematic result that when *A* is information and *not-A* is information, *A & not-A* is both information and not information. These are just some technical considerations for a quantitative account of semantic information.

Continuing on, there are other potential problems for any FI. Floridi argues that if any type of well-formed, meaningful data counts as information then we miss out on one important sense in which information can be destroyed: 'information becomes semantically indestructible and the informative content of a repository can decrease only by physical and syntactical manipulation of data' (Floridi, 2005b). Take a tourist information pamphlet. If it were to be shredded, then the information contained in the pamphlet would be destroyed by physical manipulation – the bearer of the information, the paper, is physically destroyed. If the text of the pamphlet were to be jumbled up and randomly rearranged, then

the information contained in the pamphlet would be destroyed by syntactical manipulation – the sentences that convey the information would be jumbled up.

However, it would also seem that information can be semantically destroyed. Consider, for example, the changing of a datum in a database, and how this change can affect the alethic value of another datum. If, on Tuesday January 1 2013 it is actually raining and a database contains the datum "today is Tuesday January 1 2013" and "today it is raining", both pieces of data count as information. If on Wednesday January 2 2013 it is not raining and the database contains the datum "today is Wednesday January 2 2013" whilst still holding the unrevised datum "today it is raining", then the latter is no longer information, and if nothing else changes there is an information decrease. The point here is that collections of data/information should be sensitive to factual changes, with the possibility of semantic decrease as well as increase, rather than just any piece of data indiscriminately being considered information.

Some good points can be taken from this argument. If information has to be true, this gives information a special significance, which it would not otherwise have. Without this qualification, information becomes synonymous with data, a synonymy arguably to be avoided in the case of semantic information. Semantic information is a dynamic quantitative and qualitative phenomenon and these aspects of information should be decreasable as well as increasable. A full accommodation of this fact is made possible when information is treated as true semantic content.

An attempt to account for this semantic loss whilst not accepting VT is given by Scarantino and Piccinini (2010). Under their approach the notion of semantic loss can be accommodated if treated as a *qualitative* rather than *quantitative* phenomenon. They give an example where all the true propositions in a chemistry manuscript are transformed into their negations. According to their approach, such a situation would involve loss of the original information, in that 'the information-carrying vehicles in the repository no longer carry the same information they used to carry' (Scarantino & Piccinini, 2010, p. 322). Although there would be the same amount of new information, there would also be a qualitative semantic loss of information. Also, the resulting information would be of a lower *epistemic value*: negating a true proposition causes information loss by semantic means since false information is epistemically inferior to true information. As they sum up:

> rejecting VTNN [the veridicality thesis for semantic information] is compatible with accounting for information loss "by semantic means" in the two senses – the qualitative and epistemic-value senses – that matter most for epistemic purposes. Moreover, our distinctions allow us to neatly distinguish between physical and syntactic information loss on the one hand and semantic information loss on the other. In the first two cases, information is destroyed but not replaced with any new nonnatural [semantic] information. There is information loss in the quantitative sense, in the qualitative sense, and in the epistemic-value sense. In the third case, information is destroyed and replaced with new (false, and thus epistemically inferior) nonnatural information. There is information loss in the qualitative sense and in the epistemic-value sense, though not in the quantitative sense.
> (Scarantino & Piccinini, 2010, p. 323)

It would seem then that we are once again at a stalemate; there are those such as Scarantino and Piccinini who think that misinformation is just an inferior kind of information, and those such as Floridi and Grice who think that misinformation is not an inferior kind of information, it is just not information. Depending on the position, information loss by semantic manipulation can be characterised as quantitative or qualitative. For those adhering to VT, quantitative semantic loss occurs when true

semantic content is replaced by false content or when certain semantic content is logically weakened, for example, when a true conjunction is replaced by its corresponding disjunction. Qualitative semantic loss would occur when one truth is replaced by another truth which is less valuable in some sense. For those rejecting VT, quantitative semantic loss occurs only when semantic content is logically weakened. On the other hand, qualitative semantic loss would occur when one truth is replaced by another truth that is less valuable in some sense, or when true content is replaced by false content.

7.5 Final assessment

As can be seen, both the acceptance and rejection of VT have their pros and cons. One point that the reader might appreciate by now is that although "information" is a flexible term, this flexibility is not unbounded. Caution must be exercised in its employment lest it become an indistinct "wildcard word" and any appeal to the notion of information should be justified through argument and application.

Let us close with a summary of key points for each of the two sides and ways that they counter use of the term "information" by the opposing side.

Those who advocate alethic neutrality and oppose VT can say that the truth condition it places on information is unduly restrictive and too much is lost if information requires truth. We already use the term "information" to refer to semantic content in general irrespective of its truth value and get along just fine, so why modify things? Also, it is information as semantic content in general (and not just true semantic content) that is processed by, and fuels the activities and actions of, semantic agents.

One consequence of this position is a certain asymmetry. Whereas the terms "misinform" and "misinformation" require falsity, the terms "inform" and "information" require neither truth nor falsity. Thus "information" and "misinformation" would not be strictly antonymous. If p is true, then it is information but not misinformation. But if p is false, then it is both information and misinformation.

This does seem counterintuitive. As Fox comments, 'it still seems that a claim to the effect that X informs Y that P has about it the suggestion that P is true, rather than being indifferent about the truth of P, as we might expect' (Fox, 1983, p. 159). He accounts for this counterintuitive consequence of rejecting VT by appealing to what are known as Grice's maxims of conversation. Paul Grice proposed four conversational principles that people observe in order to communicate cooperatively and effectively with each other (Grice, 1975). It is the first and fourth of these that Fox appeals to. The first is the maxim of quantity, which says that a contribution to conversation should be as informative as required but not more informative than is required. The fourth is the maxim of manner, which says that a contribution should avoid obscurity, avoid ambiguity, be brief (avoid verbosity) and be orderly.

As Fox puts it, 'in situations in which one is inclined to state that X has told Y that P, it is generally, or at least often, to the point to indicate as well whether P is true or false. If P is believed to be false, then in accordance with [the first and fourth maxims], it is most appropriate to use 'misinform', since this carries with it the proper aspersions regarding the truth of P.' (Fox, 1983, pp. 159-160). If Bob were to use "inform/information" instead of "misinform/misinformation" when Bob believed P to be false, then Bob would either violate the first maxim, because he was not being as informative as he could be and as is required, or he would be obliged to further state that P is false, which would violate the fourth maxim because he was not being as brief as he could be (misinformation is more efficient than further stating that P is false). It thus follows that in cases where the speaker believes P to be true, they use

"inform/information". This for Fox explains why "information" is by default generally associated with truth and misinformation requires falsity, despite a definition of "information" that does not strictly entail truth.

For those who advocate VT, we have seen that there is a good collection of principled and pragmatic reasons for maintaining a definition of semantic information that requires truth. On top of this, one might appeal to the terminological benefits involved. VT can be associated with the following conceptual/terminological hierarchy: Data -> Semantic Content -> Information (True) | Misinformation (False). The term information here unambiguously implies one alethic value and it is important that the category of true semantic content be given a direct term. Furthermore, a terminology whereby false semantic content has a direct term, but true semantic content does not, would be undesirably asymmetric.

Whilst people can loosely use the term information, and false information is not a contradictory or nonsensical term, under VT its use can be accounted for and does not mandate the classification of misinformation as a genuine type of information. A good guide to explain and accommodate usage of non-veridical conceptions of semantic information can be found in the following quote:

> [In other cases, people are] talking about information in a non-semantic sense; some other times, they may just be using a familiar synecdoche, in which the part (semantic information) stands for the whole (semantic information and misinformation), as when we speak in logic of the truth-value of a formula, really meaning its truth or falsehood. Often, they are using information as synonymous for data, or representations, or contents, or signals, or messages, or neurophysiologic patterns, depending on the context, without any loss of clarity or precision. (Floridi, 2010c, p. 406.)

When those who endorse VT speak of semantic information in the sense that it is a valuable concept to be used in the philosophy of information, they are talking about meaningful data that is also *truthful*.

7.6 Exercises

1. Is it better to characterise information loss by semantic means as quantitative or qualitative?

2. Are the reasons for FI given above good enough to defend FI successfully or can they be adequately addressed by the VT advocate?

3. Is Fox's explanation of the association between information and truth satisfactory?

4. Can you think of any further reasons to support or reject VT?

7.7 Further reading

Scarantino and Piccinini (2010), Sequoiah-Grayson (2007), Dretske (1981), Floridi (2005b), MacKay (1969), Peirce (1977), Wilson and Foglia (2011).

Chapter

8

8. KNOWLEDGE

How can information be used to define/explain knowledge?

8.1 Introduction

You are driving your car and look at the speedometer. It indicates 80 km/h and as a result you believe that the car is travelling at 80 km/h. As it happens, the car is travelling at 80 km/h but unknown to you the car's speedometer has stopped functioning and is stuck on 80 km/h. Does your true belief that you are going 80 km/h count as knowledge?

Information and knowledge are commonly associated with each other: colloquially and in dictionaries, the two terms are often treated as synonymous. Within philosophy however, information-theoretic or informational epistemology goes beyond this casual, colloquial association. It involves the development of specialised accounts of information and attempts to define knowledge with such accounts, to show how information causes or leads to knowledge. From this perspective, 'information is a commodity that, given the right recipient, is capable of yielding knowledge' (Dretske, 1981, p. 47).

8.2 Some background

Epistemology is the branch of philosophy that studies knowledge. One of the main tasks of epistemology has been the analysis of propositional knowledge. Propositional knowledge is knowledge of the form "*S* knows that *p*", where *S* stands for the knower and *p* stands for the proposition that is known. An example of such knowledge is given by the statement "Alice knows that Berlin is the capital of Germany", or "Bob knows that there is beer in the fridge". A central task of epistemology concerns finding a definition of propositional knowledge: what are the necessary and sufficient conditions for propositional knowledge?

Traditionally, knowledge has been defined as justified true belief. According to this justified true belief (JTB) analysis of knowledge, *S* knows that *p* if and only if:

1. *p* is true;
2. *S* believes *p; and*
3. *S* is justified in believing *p*.

Something like this definition can be found discussed as far back as Plato's *Theaetetus*. To see how this definition works, we will look at each of its three parts separately. The truth and belief components are straightforward. Firstly, if a proposition is known, then it is true; false propositions cannot be known. For example, Alice can know that Berlin is the capital of Germany but she cannot know that Paris is the capital of Germany (though she can falsely believe that Paris is the capital of Germany and she can think that she knows that Paris is the capital Germany if she is unaware that her belief is false). Bob can know that there is beer in the fridge, but he cannot know that fridges fly. Secondly, if a proposition is known by a subject, then the subject must believe that proposition. It might be true that Bob is standing in front of a fridge containing beer, but unless he comes to form this belief, Bob cannot know this fact. Finally, although an account of the notion of justification is not straightforward, the basic idea in incorporating this requirement is that knowledge involves something beyond mere true belief, and a condition is required that excludes true beliefs which are not formed by a methodical or responsible process and so just *happen* to be true. If Bob guesses that there are four bottles of beer in the fridge then although his belief happens to be true we might say that it was lucky and is not strong enough to constitute knowledge. If, on the other hand, he formed his belief after opening the fridge and counting the number of bottles on the shelf then his belief is justified and is a case of knowledge.

As mentioned, this JTB analysis of knowledge was traditionally widely accepted, and it already involved the view that true but lucky beliefs cannot count as knowledge. The second half of the twentieth century saw a conclusive challenge to the JTB analysis which prompted a revision in epistemological theorising and spawned a wave of new work on defining propositional knowledge. In 1963 Edmund Gettier published a short landmark paper titled 'Is Justified True Belief Knowledge?' (Gettier, 1963), in which he refuted the JTB account of knowledge by providing a couple of simple examples showing that there are clearly cases of justified true belief that are not knowledge. Here is one of those examples. It is supposed that a person Smith has strong evidence for the following proposition:

(A) Jones is the man who will get the job, and Jones has ten coins in his pocket.

Say, as Gettier does, that this evidence consists of the president of the company telling Smith that Jones would get the job, and Smith having counted the coins in Jones's pocket ten minutes ago.

Now, (A) entails the following:

(B) The man who will get the job has ten coins in his pocket.

Based on this evidence and reasoning, Smith comes to believe (B). As it turns out, (B) is true: the man who will get the job does have ten coins in his pocket. So Smith has the justified true belief that (B).

Now imagine the following is the case. Smith, who is unaware that he *also* has ten coins in his pocket, actually gets the job. Suppose that the boss made a mistake or changed his mind about who was getting the job. Although Smith's true belief is justified, it does not appear to be a genuine case of knowledge. Smith's true belief has been "Gettierised". This is when, in general, a justified true belief is not knowledge because it is lucky, or the truth of the belief is not connected with the justification/evidence in the right way.

Interestingly, another Gettier-style example of justified true belief that is not knowledge can be found in the pre-Gettier writings of Bertrand Russell. His example, which will be used later on, involves someone

who consults a clock at time *x*. The clock displays time *x* and as a result the person comes to form the justified true belief that the time is *x* (since looking at a clock is a justified way of telling the time and the time that the clock tells is in fact the correct time). The twist is that unbeknown to this person, the clock broke down at time *x* on the previous day. Given this, it seems right to say that the person does not know that the time is *x*.[24]

Whilst the extent to which the subjects in these examples are justified can be debated, it is fair to say that these examples suffice to show that truth, belief and justification are not sufficient conditions for knowledge. In Gettier's example the justified true belief comes about as the result of a deduction that uses a justified false belief, namely the false belief that "Jones will get the job". This led some early responses to Gettier to conclude that the definition of knowledge could be easily adjusted, so that knowledge was justified true belief that depends on no false premises. This solution did not settle the matter, however, as more general Gettier-style problems were then constructed in which the justified true belief does not result from using a chain of reasoning from a justified false belief.[25]

Much work has been done in epistemology since Gettier's paper in order to try to find an adequate definition of propositional knowledge that can, amongst other things, deal with Gettierisation. The common goal is to search for what needs to be added to true belief in order to get knowledge. There is a variety of proposals concerning just what this addition might be and there can be more than one way to flesh out a proposal.[26]

8.3 Floridi's attack on justified true belief

Despite some of this work, it is argued by Floridi (2004a) that an account of propositional knowledge which tries to salvage the JTB account is doomed. This is because the Gettier problem is not solvable by any attempt to define knowledge that *p* by adding a property (or properties) X to true belief, where having property X does not *entail* that the belief that *p* is true. Regarding the JTB account, that X is justification. Although justification makes it likely that a resulting belief is true, it does not entail or guarantee its truth. In fact our justification methods are fallible and we can have justified false beliefs.

Floridi argues that a solution to the Gettier problem for JTB requires *coordination* between the truth of P and the justification for P. Such coordination would *ensure* that we can't have one without the other (truth and justification can't be accidentally coordinated), and then there is no possibility of a Gettier problem.

A natural response to this realization is to try to add a fourth condition to the justified true belief account that says "the truth of *p* and the justification that *p* are coordinated". However, adding such a condition is tantamount to adding a condition that says "Gettier problems are impossible". This hardly solves the original problem of rescuing the justified true belief account of knowledge from Gettier problems. It gives only "justified true belief in the absence of Gettier problems" is knowledge.

Floridi argues that this shows us something rather important about the justified true belief account of knowledge. We cannot just add a condition stipulating coordination between the truth of *p* and justification for believing *p*. So we can see that what the justified true belief account is seeking is an account whereby the truth

[24] Russell's passage can be found in his book (Russell, 1948), in Section D of the chapter 'Fact, Belief, Truth and Knowledge' (Chapter 11).
[25] See Hetherington (2005) for an article on Gettier Problems.
[26] See Steup (2006), Fieser and Dowden (2007) for some literature on the matter.

of *p* and the justification for *p* remain logically and empirically independent, but that nevertheless they cannot – ever – be accidentally coordinated.

Floridi says that this is not possible. Either we link truth and justification empirically (or logically), or we accept that various things follow. If truth and justification are empirically independent, which we are committed to if we accept that justification of a belief does not guarantee its truth, we will sometimes have one without the other. And we will sometimes have both truth and justification, but in a way in which the coordination between them is accidental – i.e. we will sometimes have Gettier cases.

Floridi concludes that Gettier has shown that the justified true belief account is *irreparable in principle*. We should stop trying to fix it, and look for a different account.

As alluded to in Chapter 7, a conception of information where information must be true can help here. If the additional property were to be something like the property of being based on information, then since this property entails that the relevant belief is true (because information is true), such an analysis would deal with this issue. This type of solution does not aim to fix the JTB account; it eschews the traditional notion of epistemological justification altogether. Dretske and Floridi are two philosophers who have used information to construct accounts of knowledge that, amongst other things, overcome the Gettier problem.

Dretske's information-theoretic epistemology has been quite influential. It involves a definition of information that encapsulates truth and then defines knowledge as information-caused belief. In this way, S's belief having the property of being information-caused entails that it is true.

Floridi offers an account of knowledge that differs from standard approaches to defining knowledge, abandoning them in favour of an informational approach that is non-doxastic; that is, it is not based on belief. In short, information is true semantic content and knowledge is information that has been correctly accounted for.

So under both of these accounts information is a fundamental precursor to knowledge. Truth is embedded into knowledge because knowledge is based on information, which itself is also veridical (i.e. it entails truth). We shall now look at these accounts in more detail.

8.4 Dretske's information-theoretic epistemology

Dretske's approach to information is discussed in detail in section 6.4, which can be reviewed if necessary. Here, we deal with Dretske's approach to knowledge.

8.4a Dretske on knowledge

With his definition of information in hand, Dretske gives the following definition of knowledge:

> *K* knows that *s* is *F* = *K*'s belief that *s* is *F* is caused (or causally sustained) by the information that *s* is *F*.

The account given is restricted to perceptual knowledge of contingent states of affairs (i.e. states of affairs having an informational measure of something greater than zero given Dretske's definition of

information) and deals with only knowledge *de re*, not knowledge *de dicto*. Roughly, knowledge *de re* is knowledge of a particular thing, whereas knowledge *de dicto* is knowledge that something is the case. For example, if one sees the flag of Switzerland, without knowing that it is the flag of Switzerland, one can come to know *de re* that the flag of Switzerland has a cross without knowing it *de dicto*.

Since Gettier cases have exposed the shortcomings of the justified true belief account of knowledge, an account that is not based on this tripartite foundation was, and still is, sought. With his information-theoretic account of knowledge Dretske attempts to do away with, or at least reduce the importance of, the "philosopher's usual bag of tricks" such as justification, reasons and evidence (Dretske, 1983) in order to provide a more viable account of perceptual propositional knowledge. Gettier-like difficulties arise for accounts that make knowledge a product of some justificatory relationship between the agent and what the agent believes and this could relate one to something false. The information-theoretic account, on the other hand, relies on certain things external to the knower working out, which deals with the problem because although an appropriate justificatory relationship might lead to a false belief, an information-caused belief cannot be false. Also, such an approach allows for a more general, less anthropomorphic account of knowledge that could provide a way to attribute knowledge to infants, animals and even artificially intelligent agents, without having to suppose that they are capable of sophisticated human operations such as introspection, reasoning and justifying.

Given Dretske's definitions of information and knowledge, it might seem that his analysis is viciously circular. Recall that in his definition of information k stands for the background knowledge the epistemic agent already has about the source. If knowledge is analysed in terms of information, and information is in part analysed in terms of knowledge, a seeming problem is that the reference to k prevents the definition of knowledge from getting off the ground; one would already have to have a definition for knowledge in order to apply it.

Dretske argues that this is only an apparent vicious circularity and that his definition of knowledge is intended to be recursive in nature. Although the definition is circular, it is not viciously so, and recursive application of the definition will in all cases terminate. To make this clearer, consider the following example, an adaptation of one of Dretske's own.

> **Shell Game:** There are three shells and a peanut is located under one of them. You investigate shell 1 and find that the peanut is not under it. Given what you already know, two possibilities remain. You subsequently investigate shell 2 to discover that it also doesn't contain the peanut. At this point you deduce that the peanut is under shell 3.

For you, the observation of shell 2 carries the information that the peanut is under shell 3 given what you already know about the situation. Now, let us go through this example backwards. You know that the peanut is under shell 3 because you receive the information that it is not under shell 2 by inspecting shell 2 and you already know that it is not under shell 1. How did you come to know that it is not under shell 1? Well, you came to know that it is not under shell 1 purely through visual signals, without any background knowledge. It is at this stage that the analysis must come to an end and no appeal to background knowledge is required.

With this out of the way, the next question to ask is how can beliefs be caused by information? Dretske explains this with a simple example. Suppose that you are waiting for your friend at home and that the two of you have pre-established that the action signalling their arrival will be three quick knocks on the

door, followed by a pause, followed by another quick three knocks. It is that particular signal, that particular rhythmic pattern, which constitutes the information that your friend has arrived. Things like the knock having a certain pitch or amplitude are irrelevant. When it is this particular rhythmic pattern of knocks that causes you to believe that your friend has arrived, then we can say that the information that your friend has arrived causes you to believe he has arrived. The knock might have other incidental consequences; a loud knock might frighten away a nearby mouse or cause the window to vibrate, but what causes these things is not the information, because these things would have occurred with any loud rhythmic pattern.

Also, the causally sustained qualification in the definition of knowledge is an important one. Imagine the football team you follow play a game of which you do not yet know the result. You take a guess and come to form the mere true belief that they won the match. The following day you consult the scores in the newspaper which gives you the information that they won and your mere true belief that the team won then becomes knowledge. Although this piece of knowledge was not initially caused by the information, it was *causally sustained*.

Before closing this section, let us look at some philosophically interesting consequences of Dretske's account that have provoked much discussion. Dretske's account of knowledge is a form of epistemological *externalism*, according to which knowledge depends on factors external to the knowing agent. As he puts it:

> Externalism is the name for an epistemological view that maintains that some of the conditions required to know that *P* may be, and often are, completely beyond the ken of the knower. ... The idea is that the information required to know can be obtained from a signal without having to know that the signal from which you obtain this information actually carries it.
> (Dretske, 2008, p. 39)

So if you look at a clock and it is working properly, then you can come to know the time, irrespective of whether you can be sure that it is working properly. You don't need to stick around for another minute to make sure that the clock moves, or open up and inspect the clock to verify that it is working. As long as it is working the clock is giving you information, from which you can form knowledge.

Thus such an informational epistemology defines knowledge, while allowing us to fail to know when we have knowledge. As a consequence a seemingly plausible general principle in epistemology known as the "KK principle" or "positive introspection" fails. According to the KK principle, if one knows that *p*, then one knows that one knows that *p* (Hemp, 2006). But since with Dretske's account the conditions required for information flow and knowledge do not themselves need to be known, it follows that one can acquire knowledge without knowing that one has done so. In order for Alice to know that she knows that *p*, she would need the information that the signal carrying the information that *p* is carrying the information that *p*. In the clock scenario, Alice has the information that the time is such-and-such, but without getting some meta-information about the clock, that it actually is working, she doesn't know that she knows that the time is such-and-such.

Another related consequence of Dretske's account is that knowledge closure fails (see (Collins, 2006; Luper, 2010) for more information on epistemic closure). There is a variety of ways to construct a precise definition of knowledge closure, but the following broad formulation captures the gist of it:

> If S knows that *p*, and comes to believe that *q* by correctly deducing it from her belief that *p*, then S knows that *q*.
> (Collins, 2006)

The closure principle is quite intuitive. Take the following example: Bob knows that lunchtime is 1pm. From this he knows that if it is 12:30pm, then lunch is half an hour away. By looking at a functioning clock Bob comes to know that it is 12:30pm. Bob correctly deduces and therefore knows from all of this that lunch is half an hour away.

However, as Dretske would argue, there are some propositions that you cannot come to know like this. Dretske's zebra-mule example demonstrates this point (Dretske, 1970):

> **Zebra:** Alice is at the zoo and goes to the enclosure marked "zebras", where there is actually a zebra. She sees the zebra and comes to form the belief that the creature before her is a zebra.

According to Dretske, the visual signal of a zebra provides Alice with the information that there is a zebra and so Alice knows that there is a zebra. Alice also knows that something being a zebra entails that it is not a cleverly painted mule disguised to look like a zebra. If she correctly deduces and comes to believe that the animal before her is not a disguised mule, does this count as knowledge? Dretske is going to say no, because the visual zebra-like signal that Alice receives is not enough to distinguish between an actual zebra and such a fake zebra; for all Alice's information it could be a mule. But if this is the case, how can Alice know that it is a zebra in the first place?

The details are not straightforward, but basically Dretske's idea of *relevant alternatives* comes into play here. Assessment of the zebra proposition is made against a certain set of relevant alternative scenarios. Amongst these relevant alternatives are standard zoo scenarios such as one where there is no animal in the enclosure, one where there is an ostrich in the enclosure, one where there is a giraffe in the enclosure, etc. The alternative in which a disguised mule is in the enclosure is far-flung and not relevant. Since the visual zebra-like signal suffices to rule out all the relevant alternatives, it carries the information that there is a zebra. In other words, in all the relevant alternatives, we would not come to believe that there is a zebra were it not for the visual zebra-like signal that we receive. Assessment of the not-mule proposition does, however, have amongst its relevant alternatives ones in which there is a disguised mule; whilst "disguised mule" is not relevant to "zebra", it is relevant to "mule". Since the zebra-like visual signal cannot rule out these relevant alternatives, it does not carry the information that the animal is not a disguised mule.

As another example to help convey this idea in a different way, suppose that you read a book which provides you with the information that Bertrand Russell was born in 1872. As a result you come to know this fact. You also know analytically that the occurrence of this event implies that the world is older than a few minutes and that the world did not spring into being five minutes ago complete with a simulated history. But the book cannot provide the information that the world did not come into being five minutes ago. The book provides information about the historical event given that the past is real. It doesn't provide the information that the past is real. In this way, you can have the right information to know that Russell was born in 1872 (given that the past is real and it did actually occur), without having the information to know that you are not living in a world that sprang into being five minutes ago.

Dretske's rejection of closure offers a way around skepticism. A standard skeptical argument in epistemology goes something like this. You start off by claiming that you do not know that it is *not* the case that some skeptical hypothesis *SKEP* is true (e.g. that it is actually a disguised mule in the zoo; that the world sprang into being five minutes ago, etc.). This is fair enough; for example, there is nothing in your experience to distinguish being in the actual world to being a world five minutes old with a simulated history. Secondly, you make the claim that you know some ordinary proposition *P* that implies not-*SKEP*. For example, let *P* stand for "Russell was born in 1872" and let *SKEP* stand for "the world sprang into being five minutes ago". You then have the following reasoning:

1. not know not-*SKEP*
2. Know P
3. Know (P then not-*SKEP*)

It follows from 2 and 3 that

4. Know not-*SKEP*

But this means that you both know and do not know *SKEP*; 1 and 4 contradict each other. Given that contradictions are unacceptable, the skeptical challenge claims that this shows we cannot come to know ordinary propositions such as *P* in the first place. Dretske gets around this by denying closure, that is, by denying that 4 can be inferred from 2 and 3.

Thus for Dretske information and knowledge are not closed under informed/known entailment. This rejection of closure is a somewhat radical idea and has proved quite controversial. It is important to note, though, that it is not an outright rejection of the possibility of gaining knowledge by deducing from what we already know. In most cases closure does hold. For example, if you see a zebra at the zoo, then since the visual signal can distinguish between zebra and tiger, you can use your knowledge that the animal in the enclosure is a zebra to deduce that it is not a tiger. It is only when we go from knowledge of ordinary propositions to knowledge of skeptical ones like disguised mule or simulated history that closure fails.

There are other ways to employ the notions of information and relevant alternatives without denying closure. One option is to adopt some form of contextualism (Black, 2006; Rysiew, 2011). Contextualism is a name given to a group of theories that were initially inspired by Dretske's relevant alternatives theory and the desire to preserve the principle of epistemic closure as well as deal with the problem of scepticism. Put generally, epistemological contextualism is the view that the truth-value of an attribution of knowledge varies according to some context. Attributions of "knows that" are made relative to contexts, so that they might be true in some more relaxed contexts and false in some stronger contexts. Applying contextualism to an account of knowledge involves identifying some parameter of that account and varying it according to context. In the case of information-theoretic epistemology, this parameter can be the set of relevant alternatives. You can shift contexts, but each knowledge statement being considered at one time is judged in the same context. Reusing the zebra example above, the contextualisation works something like this. In a lower standards context, scenarios in which there is a disguised mule in the enclosure are ruled out as being irrelevant. In this context, Alice both knows that the animal is a zebra and that it is not a disguised mule given closure. In a higher standards context, alternatives in which there is a disguised mule are considered relevant. In this context, Alice does not know that there is a zebra before her and so the deduction of closure doesn't come into play.

8.4b Applying Dretske's account of knowledge to some cases

Dretske's account can easily deal with the above example from Gettier, where Smith believes, correctly and with justification, that the man who will get the job has ten coins in his pocket, although Smith falsely believes that *Jones* will get the job. On Dretske's account, Smith has not received the information that the man who will get the job has ten coins in his pocket. Clearly the boss's word is not a source of information and one of the premises used in the deduction is false. There are two ways of looking at this. The first is to say that if some piece of non-information is used in deducing some truth then that truth is not information. The second is to say that the signal consisting of the boss's word *and* counting Jones' ten coins does not carry the information that the person who will get the job has ten coins in his pocket.

As can be gathered by now, this account also easily explains why knowledge is not present when one forms a justified true belief that the time is such-and-such based on looking at a broken clock that happens to be stuck on the right time. The clock's signal does not carry the information that the time is such-and-such because its signal is not reliably correlated with the time when the clock is broken.

Other examples of epistemic luck further demonstrate the workings of this information-theoretic epistemology. Consider Alice who, upon looking into a sheep's field, sees something that looks like a sheep and forms the true belief that there is a sheep in the field. Unfortunately for Alice, however, what she is looking at is not a sheep but a sheepdog that looks like a sheep. Nevertheless, her belief is true since there is a sheep in the field, hidden from view behind the dog. This example is different from the ones provided by Gettier, but nonetheless involves an element of luck that precludes knowledge.

In terms of information-theoretic epistemology, we can say that the sheep-like visual signal does not carry the information that there is a sheep in the field. This is because there are relevant alternatives in which the dog's presence results in a sheep-like visual signal even though there is no sheep present. Since this signal causes the belief in question, the belief that there is a sheep is not caused by the information that there is a sheep.

Next take the barn-facades example, which Alvin Goldman uses, but attributes to Carl Ginet (Goldman, 1976). Suppose Bob often drives through an area in which there are many fake barn facades, although he is not aware they are fakes. Amongst the fakes is one real barn. Usually when driving through this area Bob will form a false belief that there is a barn in front of him when he is in fact looking at a barn facade. Nevertheless, since Bob has no reason to suspect that he is the victim of such a setup, his beliefs are justified. Now suppose further that on one of those occasions when he believes there is a barn in front of him Bob happens to be looking at the one and only real barn in the area. This time his belief is justified and true, but since he could very well have been duped as he often is, Bob can be considered lucky that he was in front of the one real barn at the time. Thus it is fair to say that his belief is not an instance of knowledge.[27]

For this case Dretskean information-theoretic epistemology would say that Bob's true belief that there is a barn in front of him is not based on the information that there is a barn in front of him, since in this context the belief-causing visual signal of a barn does not carry the information that there is a barn. Given the context, there are relevant alternatives where the visual signal of a barn results from a barn facade

[27] Note that some epistemologists do not regard the fake barns case as being a genuine Gettier case. There is a touch of vagueness in what counts as a Gettier case. See Hetherington (2005) for more.

being in front of Bob, so it is possible amongst the relevant alternatives that "barn-like signal" and "not-barn".

These last two examples in particular demonstrate a benefit of such an externalist epistemology. If the conditions external to the knowing subject are not right, the absence of the right conditions can easily be used to account for the absence of knowledge.

There is a range of purported counterexamples and objections to Dretske's account of knowledge. The usual strategy is to provide an example where it is correct to say that the subject does not know that *p* and claim that Dretske's account is committed to saying that the subject does know that *p*. The objections generally fail because the cases they give would not actually be classed as knowledge given a genuine and correct application of Dretske's account, but some objections are better than others and some compel us to reconsider Dretske's definition. See Doyle (1985), Dretske (1983), Adams (2005) for some critical discussion of Dretske's account.

8.5 Floridi's informational epistemology

Recall that at the start of this chapter we came across the definition of knowledge as justified true belief (JTB). According to this definition, Bob knows that there is beer in the fridge if and only if:

1. It is true that there is beer in the fridge;
2. He believes that there is beer in the fridge; and
3. He is justified in believing that there is beer in the fridge.

For example, suppose that Bob bought a box of beer during the day and put it in the fridge upon returning home. That night, knowing that there is beer in the fridge, he goes to the fridge and gets a bottle. He knows that there is beer in the fridge because he believes this fact and his belief is justified given the day's events.

Or as another example, Einstein knew that the General Theory of Relativity is a good theory if and only if:

1. It is true that the General Theory of Relativity is a good theory;
2. Einstein believed that the General Theory of Relativity is a good theory; and
3. Einstein was justified in believing that the General Theory of Relativity is a good theory.

Well, Einstein worked out the General Theory of Relativity. He also knew about its vast empirical success, beginning with the light-bending experiments and going from there. Therefore his true belief is justified and counts as knowledge.

Despite these plausible explanations, we saw earlier how the Gettier problem created problems for the justified true belief account of knowledge, so much so that Floridi argues that the JTB account cannot be mended, and should be abandoned. Briefly, this was because the account seeks both to keep justification and truth logically and empirically independent, and also ensure that they cannot be accidentally coordinated, which Floridi says is impossible.

8.5a The network theory of account

Recall Floridi's amendment of the General Definition of Information (GDI) from section 7.2. According to the GDI, something is an instance of semantic information if and only if:

1. that something consists of one or more data,
2. the data are well formed,
3. the well-formed data are meaningful, and
4. the well-formed meaningful data are truthful.

Criterion 4 is Floridi's addition and results in what is known as the veridicality thesis, which was covered in Chapter 7. So, according to Floridi, when a person holds a piece of semantic information, that person holds something that is meaningful and truthful. However, Floridi holds that although Bob's belief that "There is beer in the fridge" might well be both meaningful and truthful, this is not yet enough for it to count as knowledge.

Floridi's network theory of account (Floridi, 2012b) tells us what the extra thing is that a person with semantic information needs in order to have knowledge. It can be stated briefly: knowledge that *p* is correctly accounted for semantic information that *p*. To be accounted for, semantic information needs to be embedded in a network of questions and answers that sufficiently answers certain "How come?" questions that the semantic information raises.

Floridi uses network theory to give a formal account of networks of relations among information. However, we will examine the basic idea here without going into network theory. Floridi takes a major step beyond many standard epistemological approaches to hold that knowledge is essentially integrated. We don't "know that *p*" all by itself! Whenever we know something, we know a bunch of related things. Bob can't just "know that there is beer in the fridge". He also has to know what beer is, what drinks are, what fridges are, and what they are for. He wouldn't get far without knowing what kitchens are, and why we have them in houses. If you pause to consider that for a moment, you will see that this also involves a great deal of knowledge of human beings, why they need shelter and food, and our normal cultural practices for meeting those needs that lead to us building houses with kitchens – and putting fridges in them.

So each piece of information poses certain questions that can be correctly answered by other pieces of information. For the semantic information that *p* to be knowledge, it needs to be embedded in a network of related information that correctly and sufficiently answers certain questions it raises. This links it, as it should, to its potential role as evidence, its potential value for prediction, inferential processes, explanation and so on.

The idea of knowing agents having a network of information is the first idea. The second element is that of accounting. The network of related semantic information accounts for the semantic information that is to become knowledge. For Floridi the informational level of abstraction (LoA, see Chapter 2, and Chapter 4) is the most general way of considering and describing all things. In this sense, the world is made of information – it is all data.

The basic idea of accounting is that the semantic network, the information that a knowing agent has, is in *connection* with the data in the world. The data in the world is the source of semantic information, which has to be sensitive to the data, and stay sensitive to the worldly data. Accounted for semantic information changes, in connection with data, so that information flows correctly from the source through the network. For someone with a network that accounts for their semantic information, the world is the source of their semantic network,

their model of the world. As a result that model is generated by the world, models the world correctly, and stays correct by being sensitive to ongoing changes in the world that is the source of the information in the network. Such a person would seem to have knowledge!

Thus Floridi has a definition of knowledge according to which an epistemic agent *S* knows that *p* if and only if:

i) *p* qualifies as semantic information (it is well-formed, meaningful and truthful data);
ii) *q* accounts for *p;*
iii) *S* is informed that *p;*
iv) *S* is informed that *q;* and
v) *S* is informed that *q* accounts for *p*.

To make this clearer by going back to our examples, Bob knows that there is beer in the fridge when he has accounted for semantic information that there is beer in the fridge. That there is beer in the fridge is a piece of information that itself raises certain questions, such as how did the beer get in the fridge? It is when Bob has a network of related semantic information about the beer that can answer questions such as these that he has the knowledge that there is beer in the fridge. This network is in connection with the data source. In a scenario where Bob has just bought the beer and placed it in the fridge, his information has the right kind of relations to evidence (such as he bought the beer and put it there, and has seen it there since buying it). His information has the right kind of relations to other beliefs: his fridge is in the kitchen, it is keeping things cold. His information has the right kind of relation to his future actions, such as his action of going to the fridge to find beer, and into the kitchen to find the fridge.

Our second example about Einstein knowing that the General Theory of Relativity is a good theory is particularly interesting. It is very clear that knowing any scientific theory is good requires both having information about a *lot* of related things, and being in touch with the world – the data source – in the right way. (See also Chapter 9.) Examples from science are not commonly used in epistemology, even though scientific knowledge is arguably our most important kind of knowledge. On Floridi's view, Einstein had such knowledge if and only if he had accounted for semantic information regarding the General Theory of Relativity being good. This network of information means the theory has the right kind of relations to evidence. Einstein worked on it, and knew very well the empirical evidence and its significance. His information had the right kind of relations to other beliefs, such as information about odd empirical results and many other physical and mathematical theories. His information had the right kind of relation to his future actions, such as he will continue to work on problems, and have expectations about new empirical work, on the basis of the General Theory.

Recall that the Gettier problem arose by allowing the justification and truth of a belief to be logically and empirically independent, which allows the coordination between the justification and truth to be accidental, such as there still being beer in the fridge only because Bob and Alice's teenage daughter Carol drank the old beer and replaced it with new beer.

Floridi uses a constructive strategy to break away from the constraints that make the problem unsolvable. If an epistemic agent has a network that accounts for the information that *p*, then the information is coordinated to the source that correctly accounts for it. So there can't be accidental accounting, and so no Gettier-type counterexamples. If the world changes, due to Carol drinking the beer, Bob's network ceases to account for the information. He might be able to answer whether there is beer in the fridge correctly (there is because Carol replaced it), but his network can no longer correctly answer the relevant question concerning how it got there. His model is no longer sensitive to the world in the right way, and he no longer counts as having knowledge, on

this account, unlike on the JTB account. His former information that the beer is in the fridge because he put it there becomes misinformation.

Einstein knowing that the General Theory of Relativity is good is an excellent example of how the integration of semantic information is the right way to avoid Gettier problems. The core idea is that the "justification" for believing *p* just is all the integrated network of related information, connected to the data source. This is information about the problems with prior physical theory, like Newton's. This is information about many empirical observations such as the movements of planets and stars, and also of physical things on the earth. This is information about related physical theories such as electromagnetism and so on. But all this information, the connection to the data source, is just what makes the theory good. The truth of the claim cannot disconnect from the "justification" for the claim. One could not "accidentally" come to have this kind of network of accounted for semantic information that the General Theory of Relativity is good.

This solution might look like a cheat. Recall that Floridi says that proponents of the JTB account cannot add a fourth condition to say "and the truth of the belief and its justification are coordinated", as this is merely to say "Gettier cases are impossible". But while proponents of JTB are precluded from coordination – tying together the truth of *p* and its justification – as a solution, Floridi is allowed to seek this kind of solution. Attempting to maintain the JTB account of knowledge commits you to maintaining truth and justification as logically or empirically independent. This is why you cannot solve the problem. Floridi, however, does not accept the frame of the justified true belief account, and so is not precluded from coordinating truth and justification.

A final point to make is that Floridi's approach takes an unorthodox turn when it comes to dealing with basic perceptual knowledge and testimonial knowledge. Take an example involving a car which has a red light that flashes when the engine temperature exceeds a certain threshold. Alice is driving the car, visually perceives that the red light is flashing and thus we can say comes to know that the red light is flashing. If she knows what the light indicates, then she will probably stop the car and check the radiator. If she is unaware of the light's purpose, then she may very well continue driving. Either way, she knows that the red light is flashing. Dretske's account will explain this knowledge straightforwardly; the red light signal carries the information that the red light is flashing and this causes the belief that the red light is flashing. Likewise, other prominent accounts of propositional knowledge will accommodate this example as a case of knowledge.

As Piazza (2010) discusses, if, as the network theory of account has it, knowledge is information that has been correctly accounted for, then how can this analysis explain perceptual or testimonial knowledge? For example, it would seem that a basic perceptual experience of the light flashing is not enough to know that the light is flashing; it does not suffice to give an account and answer relevant questions about why the light is flashing. Floridi deals with this by reinterpreting perception and testimony as data providers rather than full-blown cases of knowledge (Floridi, forthcoming-c). Thus if Alice sees the red light flashing but cannot explain the context or why it is flashing, then she does not have a sufficient account to know that the red light is flashing; at best she merely holds the semantic information that the red light is flashing.

8.5b Criticisms and benefits

Floridi acknowledges that he has not answered any sceptical question. We could think we have accounted for semantic information, and yet be brains in vats. This is in accord with his stated aims for the philosophy of information. Sceptical questions have been unanswered for at least two millennia. Floridi thinks that they are not answerable, and so they are bad philosophical questions. Attempting to answer sceptical concerns is fruitless, he thinks, and so should be abandoned (see Chapter 2).[28]

A more important natural worry is whether Floridi has moved beyond a sophisticated *coherentism*. You might recall from other philosophical studies that foundationalism about knowledge involves looking for certain fundamental pieces of knowledge, and building all further knowledge on them. Descartes uses the cogito. Empiricists use sense-data, or some more modern empirical observations. Coherentists deny that this search for privileged foundations for knowledge – pieces of knowledge so strong that they can, as it were, hold up all the rest – is necessary. Indeed, they often deny that there are any such privileged pieces of information. For coherentists, you have knowledge when you have an integrated web of coherent beliefs. Clearly, Floridi's view has something in common with coherentism, although of course Floridi's network is one of information, not belief. However, Floridi has moved beyond coherentism in the idea of having the network accounted for. This means it must be in the right kind of contact with the data source – the world. We build our networks of information by repeated interaction with the world, and so gradually correct them over time.

It might also be thought that Floridi's account is circular. Floridi notes that an open problem in the philosophy of information is the information circle. This is the problem that we check information with other information, and it looks as though that is precisely what Floridi does. Naturally, Floridi's philosophy of information is committed to this - there is nothing else to check information with! But is this a serious problem? Consider in comparison the same kind of question pressed against traditional accounts of knowledge. A question sometimes asked is: do you have to *know* the justification for *p* before you know *p*? If you think about it, there is nothing stronger that we can check knowledge with except knowledge. Does this mean that knowledge is also stuck in a knowledge circle? It seems that the problem is not as serious as it first appears. If you think that this is reasonable, and that we still have knowledge at least sometimes, then you need to try to produce an account of how some pieces of information – or beliefs – amount to knowledge, while others don't. Floridi has certainly given us that.

A very good aspect of the view is that it seems to offer a much better approach to scientific knowledge than traditional epistemological approaches. It is an excellent point that generally knowledge of a single proposition, isolated from related propositions, is impossible. This is clearly true in science. So the basic starting point of the account has to be right – knowledge has to be about integration. Indeed, our most important current knowledge, scientific knowledge, is clearly highly integrated. Interestingly, the account also allows knowledge to come in degrees. On traditional accounts, knowledge is all-or-nothing: you either have it or you don't. But you can have a more or a less extensive network. And this allows two people both to know something, but one to know it better than the other. That person has a more extensive accounting network.

Note also that the network theory of account affords us a natural way to see how a community of people can hold knowledge communally. A community of people can build a network of account, and different people can

[28] See also (Floridi, 2010b, forthcoming-a) for some papers on this matter. In the former Floridi investigates the sceptical challenge from an information-theoretic perspective, contending that informational scepticism is not a problem. In the latter paper Floridi defends, contra Dretske, a principle of information closure.

hold different bits of it, so long as it is accessible to other people. Such an application of the network theory of account brings in related issues such as trust and testimony.

Interestingly, we can apply this philosophical account of knowledge to explain scientific knowledge, which is essentially the knowledge of a very large community. In scientific papers, both written and oral, and in large and small databases containing the results of experiments, different individual scientists hold some parts of the network of information. And the increasing availability of much of this online allows the network of scientific information to be built faster and disseminated further than ever before.

8.6 Exercises

1. Two suitable formulations of the principle of information closure are: (1) if *S* holds the information that *p* and the information that *p implies q*, then *S* holds the information that *q*; and (2) if *S* is informed that *p* and *p carries the information that q,* then *S* is informed that *q*. Is it plausible to suggest, as Dretske does, that this principle can fail in certain cases?

2. Suppose that your car is moving at 80 km/h. Dretske's account explains why if you look at your car's functioning speedometer showing 80 km/h you can come to know that this is its speed, whereas if the speedometer is not working and just happens to be stuck on 80 km/h, any true belief you form as a result of looking at the speedometer does not count as knowledge. Imagine the following scenario. Your speedometer works perfectly up until 100 km/h at which point it malfunctions and cannot move any further up. When your car is going 80 km/h, does the speedometer's correct signal indicating 80 km/h provide you with the right information for knowledge, or does its range of unreliability above 100 km/h affect its general information-carrying potential?

3. Can you think of any examples that are problematic for Dretske's definition of knowledge?

4. Consider the difference between a coherentist account of knowledge, which has beliefs connected only to other beliefs, and Floridi's account, which has a network of information also connected to the data that is the ultimate source of that information.

5. Do you know that the General Theory of Relativity is good? Do you know it as well as Einstein did? What is the difference?

8.7 Further reading

Dretske (1983), Bremer and Cohnitz (2004), Floridi (2012b).

Apart from Dretske's and Floridi's applications of specialised accounts of information to analyses of propositional knowledge, there are other pockets of work in philosophy on information and epistemology. Two examples are Fallis (2002), Harms (1998).

Part IV: (Information in the) Sciences

Chapter 9

9. SCIENCE

Evidence and expert knowledge

9.1 Introduction

> *'We must not forget that when radium was discovered no one knew that it would prove useful in hospitals. The work was one of pure science. And this is a proof that scientific work must not be considered from the point of view of the direct usefulness of it. It must be done for itself, for the beauty of science, and then there is always the chance that a scientific discovery may become like the radium a benefit for humanity'. Marie Curie*

We read a lot in the news, on blogs, and on social network sites about the advancements, achievements, and even the failures of science. Think about the recent experiments at CERN. Scientists thought they had found neutrinos that travel faster than light, and then they had to retract the claim. Another recent piece of news from CERN is the discovery of Higg's boson. Higg's boson, they claim, has been "found" using two different experimental set-ups. This is a milestone in the history of physics, not just because it confirms a whole theoretical apparatus, but also because it opens up new paths of research. Consider also something we are all concerned with: health. We want to be cured when we are ill and we hope that the drugs available are effective and do not produce bad side effects. Examples abound of drugs that provoked disaster – think about the famous Thalidomide scandal. And examples abound of drugs that are effective – the acid acetilsalicilic, also known as aspirin, has been used for centuries. Science often produces major news headlines.

In this chapter, we examine science, in its efforts to explain, predict and control our world, via *modelling* the world. Traditional philosophy of science, coming out of the work of the logical positivists, focused on the language and logic of scientific methodology. We look at what light the Philosophy of Information (PI) can shine on traditional philosophical issues concerning science as it is practiced now. First we examine how science finds things out, looking at the importance of model validation in science, linking that with Levels of Abstraction (LoAs) from Chapter 2. Then we examine what counts as evidence for the things we discover, seeing how understanding evidence as information can deepen understanding. This allows us to move on to look at what kind of knowledge science gets after model validation and gathering of evidence, seeing how the idea of knowledge as a network of accounted for information fits well with science. This last idea is developed extensively in Chapter 8, and comparing this chapter with that shows what the problem of scientific knowledge has in common with the problem of knowledge in general.

So far these are traditional issues in philosophy of science, focused on science discovering things in the world. But what about making the world? Making is a core concern of philosophy of technology. After the information revolution, PI sees us more than ever as creators of our world, creators of the infosphere. Traditional approaches in philosophy of science neglect this aspect of what science does, treating technology as if it were merely an application of science. But with the creation of advanced technology such as the large hadron collider at CERN now essential to our pursuit even of fundamental science, design and creation cannot be treated as an afterthought to science, as they are deeply embedded in the practice and success of current science. This chapter closes by looking at the philosophy of technology idea of *homo poieticus* – people as makers.

9.2 Science and reality: Model validation

What is science? This is an old, hotly-disputed philosophical question. What makes astronomy scientific, unlike astrology? Why do we believe the claims of experimental medicine more than those of homeopathy? The problem of demarcation was famously set up by Karl Popper. For Popper (1963), science is about the possibility of falsifying hypotheses. Thus while we can test – and possibly falsify – a hypothesis about the motion of a planet, it is hard to think about how we could test – and possibly falsify – the hypothesis that all Gemini are stubborn. Popper's proposal is to examine which data and which methods allow us to make decisions about the hypotheses that science formulates.

Many others in history and philosophy of science have addressed problems of scientific method. Some names of pioneer methodologists will be familiar to you: Galileo and Newton, Francis Bacon and William Whewell, John Stuart Mill. These scientists (and philosophers, for that matter) primarily wanted to understand the world around them and to "codify" procedures to do this effectively. Why would they need to "codify" such procedures? Because to explain, predict and control phenomena, especially when little background knowledge is available, is a challenging enterprise. So, if we find that one method proved successful on several occasions, we may want to mainstream it.

We shouldn't underestimate the importance of the attempts to develop methods for scientific research. In accordance with the Copernican, Darwinian and Freudian revolutions (see Chapter 2) science changed our understanding of the world, and also of ourselves. But science also had to create *science*. Science creates labs, and research groups, and large hadron colliders, and ultimately science develops, examines, and revises its own methodology as well as its conceptual baggage. While we all have some knowledge and understanding of gravitation from school, Newton had to make up an explanation for the phenomena he was observing. We are now used to understanding that individuals' histories are affected by their socioeconomic contexts, but Emile Durkheim, a pioneering methodologist in sociology, was revolutionary in trying to explain differences in suicide rates according to different situations people live in. Galileo beautifully explains his reasoning and experiments about moving objects, and he changed the way we see the surrounding world, so that many of the motions he described are now part of a generally understood folk physics. All these pioneering methodologies tried to say what relations hold between their observations, the methods they used to analyse such observations, and what conclusions they could draw from them. If our current science is so advanced, it is thanks to these scientists sharing and codifying their research.

In this chapter we will take up the idea of looking at scientific method, but in a modern context, looking at three issues that play a key role in the dynamics of science, namely model validation, as a great deal of the activity of current science involves building models, such as climate change models; evidence, which is a subject of intense debate in many sciences; and expert knowledge. We will see what light the philosophy of information

can shed on these issues. To understand what really allows scientists to confirm their claims we would need to look deeply at a specific science's state of the art, at its methods of analysis and evaluation of data. We will take a more general stance in this chapter, but there is still something we can say about the scientific method in general.

PI makes an important observation about the relation between the activity of modelling phenomena and the claims we formulate about these phenomena. While traditional philosophers of science investigated the relation between our claims (or propositions) and reality, current philosophy of science is much more interested in modelling in the practice of science and its relation to claims we make about the world. This creates room for useful interactions between philosophy of science and PI.

So let's start from modelling practices in science. There is something that is common to many different scientific enterprises: modelling phenomena to describe the world, understand it, predict what will happen, and intervene to make something happen. This is in accord with the Method of Levels of Abstraction (LoAs) used by PI, and introduced in Chapter 2.

Let us consider an example from the biomedical sciences. Oncologists and epidemiologists are interested in whether or not exposure to asbestos causes cancer, and if it does to what extent. To answer such a question we need a lot of data about people who have been exposed to asbestos and about those who developed cancer, so that we can examine whether more of those who have been exposed to asbestos developed cancer, when we examine similar age groups, socioeconomic classes, and so on. We need data about exposure and development of disease. Some current studies go even deeper and gather information about the biochemical changes that happen in the body in response to asbestos exposure and lead to cancer development. Epidemiological and biochemical models are currently being integrated in the hope of contributing to our understanding of disease and to developing cures. But whether asbestos causes cancer is also crucial in legal contexts. One case has just become famous world-wide. This is the 'Eternit trial'. Eternit was an asbestos factory active in the north-west of Italy. Epidemiological data showed a massive epidemic of asbestosis and lung cancer in the surrounding area. A trial lasted for decades, to ascertain the responsibilities of Eternit managers in failing to implement appropriate safety measures. If it is true that asbestos causes cancer, and if there were health safety measures available, then the Eternit managers are guilty of having ignored them, as negligence of such safety measures caused cancer in many workers. This is the essence of the reasoning. The decision of the judges in Turin has historical importance: they convicted the managers.

However, questions like whether asbestos causes cancer don't have a simple yes or no answer. This is unfortunate. In our daily experience we often wish or expect answers to be simple. To find out whether a claim – a proposition – is true, we simply go and find out. Is it true that it is roughly half past two now? If you don't know, you can check your watch, or the clock on your laptop. Is it true that CERN scientists found the Higg's boson? Well, let us Google it – we'll surely find some information that confirms or refutes it. Science doesn't really work in this day-to-day way, however. Consider scientific claims like this: "42% of Americans could be obese by 2030" (just Google "obesity forecast" and you'll hit plenty of these examples). How do we make such predictions? How do we find out whether such things are true?

PI is very much in accord with new ways of looking at questions like this:

> For in order to understand whether a model is correct, a scientist looks at the data set and considers whether the model can successfully reproduce and/or predict the behaviour of some aspect of the system being

> modelled (Davison (2003)). The better the model, the smaller the disagreement between its implementation and/or forecast and what happens in the observed system.
> (Floridi, 2010c, p. 364)

Scientists formulate a hypothesis (which is just a claim, or a proposition, after all, such as the prediction about growth in obesity above), then build and test a model. There are a number of things we need to know about models. First of all, the models mentioned in the quote above and used by science are "empirical models", namely models that use data coming from observations or from experiments. Scientific research is not all empirical. For instance some physicists never work in a lab and what they do is so conceptual that you can't tell the difference between their physics and the purest maths you can imagine. But let's stick to empirical research, as our concern was the relation between models and reality. Models are generally not treated as true or false. Instead, they are valid or invalid, useful or useless. Validity and usefulness don't have rigorous definitions, unfortunately. But we can try to grasp their meaning nonetheless.

Validity. A model is valid to the extent that the story being told using that model holds. Let us formulate the following model validation view. Scientists have a data set with observations on a phenomenon X to study, say, rates of asbestos and cancer. Scientists analyse the existing literature on the issue, then examine the data, formulate hypotheses such as "asbestos causes cancer", build and test a model and finally tell us whether they think the hypotheses are confirmed or not. Different considerations are involved here: what background knowledge is available, what type of analysis has been run on the data, what results came up from the tests, etc. All this narrative, namely the explanation of the work (data collection, data analysis, interpretation) that led to the result of the study contributes to making the model valid. Such practices also ensure inter-subjective control in science: other scientists, peers, can challenge results or conclusions, or ask for more details and justification, on each of the elements involved in modelling (Russo, 2011). We will come back to this point later when discussing expert knowledge.

Usefulness. Imagine you have never been to London and when you arrive at St Pancras station you go to the tourist information desk and you are given a big map of the whole of the UK. How useful would that be to travel around London? Pretty useless. You would need a map that doesn't cover the whole UK, but has street names for all areas in London, and possibly a tube and bus map, too. Then you can find your way around. Modelling works pretty much the same way. You may have heard of Genome-Wide Association Studies (GWAS). These studies collect information about genes and diseases, the people who have and don't have them, and other information such as where they live, their life habits, etc. These studies give a great descriptive picture of populations and some of their genetic characteristics. How much do we learn about gene expression or gene mutation? Very little. But for that different problem we have lab experiments in biochemistry, for instance. We build different models for different purposes. So it is extremely important to define the research question and to identify the modelling procedures to best address it. Models are not true or false, but useful or useless depending on the research question (Giere, 2006).

So, to wrap up: hypotheses formulated in science are not true in virtue of some kind of simple correspondence with facts or things that hold or are out there in the world. These hypotheses are true to the extent that the models built and tested tell us something valid and useful about the reality we are studying. This idea can be further examined by reading Chapter 8 on knowledge and Chapter 7 on truth.

9.3 Information and evidence

The implications of this view of model validation are numerous. An important one concerns evidence. Let's begin with an example. Suppose you say: "I am a very good student". How do we know that? You can then point to your marks of last year and of this year. You never failed an exam and always had high marks. Wow. This is evidence that lends support to your claim. Consider now more complicated cases, such as the scientific claims mentioned above about asbestos exposure and cancer or about the Higg's boson. You can always ask: how do we know that? What established such and such a result? When you tackle questions like these, you are providing the evidence for the claim. As you can imagine, in science it is not as easy to find evidence as it is to show your marks from last year.

Not surprisingly, "evidence" is the subject of a lively debate in philosophy and in the sciences. There is an immense literature trying to provide conceptual analyses and methodological approaches for evidence. For instance, there are probabilistic approaches to evidence, trying to spell out how evidence lends support to a hypothesis. When a hypothesis h is formulated, scientists also attach some probability p to it. After running the experiments, suppose the results are positive, so that now the probability of the hypothesis, given the results from the lab, is higher. Very recent approaches in medicine coined the name "evidence-based" because they want scientific results to be based on "scientific tools" rather than, for instance, on anecdotal information and expert opinion. A particularly important movement is "Evidence Based Medicine" (EBM).

We will examine issues of evidence using the discussions of the "evidence hierarchies" of EBM, and show how thinking of evidence as information can help. In the next section, we show how thinking of evidence as information that a community gathers and processes collectively can also help.

EBM movements (Sackett, Rosenberg, Gray, Haynes, & Richardson, 2007) organise possible evidence into a hierarchy, beginning with the best evidence, which is generated in (1) randomised controlled trials (RCTs) (2) observational studies (3) case reports, and at the bottom of the pyramid we find (4) expert opinion. We explain briefly what these methods are before we go on.

(1) Randomised controlled trials. Suppose you wonder whether getting drunk just before an exam would be a good idea, since when you are drunk you are more inspired and you may produce a more original essay. You talk to a scientist and s/he decides to run an RCT to establish whether your hypothesis holds. Suppose in your class there are 200 people; they will be divided in two equal groups at random. So each group will contain males and females, bad and good students, students from different nationalities and so on. Then the scientist decides that for the next five exams, one group will always get drunk before the exam, and the other will never get drunk. After the five exams the scientist collects the results and compares the two groups. If the "drunk" group performed better than the "sober" group, perhaps there is a point to drinking before exams. If the "sober" group performs better, then perhaps, as we suspect, drinking before exams is best avoided! You will have realised why we picked such an extreme case: we already have a lot of background knowledge about behaviour, about individual response to alcohol consumption etc. So you can imagine how difficult it must be to set up a RCT when there is the marketing of a new drug at stake, or the implementation of a new surgical technique. Medical people have to deal with tough problems and RCTs try to cope with such difficulties.

(2) Observational studies. Let us continue with the example above. Suppose the scientist thinks that it is unethical to force people to drink, even for the sake of science. (There are also ethics boards like the Institutional Review Boards in many universities and funding bodies to prevent scientists from doing studies

like this even if they would like to.) So s/he will try the observational route. Perhaps there is already a database with information about drinking behaviour and exam performance, or a new study can be set up asking students about their experience. This can be done in the form of questionnaires, or interviews, or focus groups etc. The point is that this data is collected without running "experiments" of the type of RCTs. Statistics will then help detect links – correlations – and possibly causal relations between drinking behaviour, exam performance, individual response, etc. It is quite a complicated story to say whether and to what extent studies of this type really can detect causal relations, but for the time being it suffices to note that it is a difficult task, albeit one that can be managed.

(3) Case reports. Let us continue with the same example. Here we can imagine that university teachers or doctors at the university medical centre came across a student who regularly took exams while drunk, and collected as much information as possible about that student. For instance, age of the student, number of exams passed or failed, in what circumstances and so on – you can list all the factors that may play a role here. A case report is a "study" on one single event of the type we are interested in. There are a number of conclusions one may draw from just one case, but it is difficult to decide how much one case can tell us about what would happen to other people. We have to judge whether the student studied might be atypical in important respects, such as in having an unusually high tolerance for alcohol.

(4) Expert opinion. At the bottom of the pyramid we find expert opinion. This includes knowledge that people "in the field" accumulated over time. It may or may not be supported by RCTs or other types of study; it is largely based on background knowledge available in the field and on personal, direct experience. For example, an expert might advise you that drunkenness makes you less alert, careless, and possibly even sleepy, or inclined to walk out of the exam and go to the cinema instead, and that these would all be bad for your exam performance.

Now you have an idea of what EBM evidence hierarchies often share, although you may have noticed that strictly speaking it is a hierarchy of methods for gathering evidence, rather than of evidence itself. The point of the evidence hierarchy is that RCTs are superior to studies that "just" collect data, which are superior to studies that analyse one case, which are superior to the opinion of experts.

PI can bring a new perspective to thinking about evidence, and hierarchies of methods of gathering evidence, because evidence presented in different ways, generated by different kinds of trials, can all be seen as information. Simply put, the evidence generated by a randomised controlled trial is information about the probability of an outcome given the presence or absence of a treatment. The evidence produced by a lab experiment is information about biochemical processes. The evidence produced by expert opinion is information about the person's experience of the given topic. There is some exchange of information that occurs between individuals who apply methods to generate evidence and individuals that evaluate the generation of evidence. We will also see below some characteristics of this "exchange of information".

We now try to make this idea a bit more formal. Here, Floridi's version of the general definition of information (GDI) can be useful (see also section 6.5 and Chapter 7): p is semantic information if (i) it consists of data, (ii) data are well-formed, (iii) well-formed data are meaningful, and (iv) the meaningful well-formed data are truthful. "True" here means 'providing true contents about the modelled system' (Floridi, 2010c, p. 201).

Let us now see how, in scientific contexts, "information" in the above sense can shed light on the notion of "evidence". More precisely, we can spell out the notion of evidence in terms of the notion of information.

Any of the studies briefly presented above and included in the evidence hierarchy generate evidence; that is, information that is needed to back up or to refute a hypothesis. Information (in the sense above) comes from the data used in the study and such data has to have the characteristics mentioned in GDI or otherwise the study will be flawed in many respects. If data are not well-formed, for instance the variables used are not comparable with each other and it is not possible to run any analysis on them. Recall that this is what crashed NASA's Mars Space Orbiter in 1999 (see Chapter 2). Or, if data are not meaningful the study will suffer from serious conceptual weaknesses – for instance, years of schooling can measure education levels, but not motivation to study. Likewise, data have to be truthful, in the sense that measurement error, in these contexts, has to be minimised. We are using an idea of truth, in accordance with the idea of model validation, which is explained in (Floridi, 2010e).

You can now see how PI can contribute to debates on EBM. It is worth noting that evidence in terms of the GDI does not fix the kind of data that is analysed. So far as the GDI is concerned, statistical information can be evidence, experimental information can be evidence, expert opinion can be evidence. Sources of evidence are treated on an equal footing. What discriminates between good and bad evidence is the information being well-formed, meaningful, and truthful.

9.4 Information and expert knowledge

We just saw that according to the EBM evidence hierarchy, expert knowledge is at the very bottom of the pyramid. Why so? In medicine, there are excellent historical reasons why knowledge of experts has been downgraded: advancements in science showed that large parts of "established" and "undisputed" expert knowledge was actually wrong. But it may be too quick to dismiss "expert knowledge" altogether for this reason. Let us see what kind of assumption is made in such a move.

Traditional epistemology (see Chapter 8) tends to focus on situations where a single person (atomic agent Alice) believes in or knows a single claim (an atomic proposition P such as "it is raining"). Then the question asked is what makes this belief knowledge. But this focus neglects important considerations, about the type of agent, the type of proposition and its source, and about the interactions between agents, particularly the multi-agent dimension of scientific knowledge.

Interestingly enough, there is also a long tradition in the philosophy of technology discussing expert knowledge. Philosophers of technology paid more attention to issues such as the nature and scope of acquisition of skills that makes experts "experts"; or the extent to which it is important how expert knowledge is "embodied", worrying about whether artificial expert systems – such as machines that make diagnoses – are expert in the same sense as doctors are. So there are plenty of fascinating issues that arise from a reflection on expert knowledge.

Again, PI is not meant to settle all these issues, but to provide new ways of looking at expert knowledge. There is a "spill-over" effect: once we adopt GDI, we then re-analyse evidence in terms of information, then the next level is to rethink expert knowledge in terms of the interactions between agents. Such interactions form a network of information (see section 8.5). Networks of information exist in labs, conferences, publications, etc. Nothing is done in isolation. Expert knowledge is a collective effort. There is a sharing of information that takes place in many ways: collaboration in producing or even criticising a paper, running experiments as a group, communicating scientific results to different audiences, and of course teaching the next generation of scientists. Moreover, the network is dynamic. Expert knowledge (i.e. the network of information) is constantly changing as data changes, or as interactions between agents alter over time, etc. In this way PI reconciles the

agent-based approach of epistemology with the communal or network-based approach of philosophy of technology. The core idea of information, and the idea of knowledge as a network of information, do this very well.

Once we see expert knowledge this way, it appears plain that it shouldn't be at the bottom of the pyramid for deciding on medical treatment, or policy decision-making. If we adopt the view above, we see that expert knowledge is a very complex phenomenon, and the traditional concept of expert knowledge does not capture its real sophistication. For example, expert knowledge is needed to set up an RCT, an observational study, and a case report. It is needed to decide what to study, what variables to observe, and for how long. It is then needed to integrate the information from these various studies into a model, and ultimately to come to conclusions about medical matters. Expert knowledge is not something separated from the rest, but inheres in the whole of science – even in very different scientific practices.

9.5 Science and *poiesis*

We have discussed traditional philosophical issues of the sciences, looking at how PI can shed new light on how scientists discover things, on evidence for those discoveries, and on the nature of scientific expert knowledge. We have seen that traditional approaches to epistemology miss the essentially collective, collaborative nature of scientific knowledge-gathering. Finally, we look at another aspect of science that philosophy of science largely misses: we don't just find out about the world; we also create it. The scientist is a maker, a *homo poieticus*, and therefore a techno-scientist, as the word is used in philosophy of technology. Here, we follow ideas of Luciano Floridi, which are influencing some other philosophers in the PI tradition (Demir, 2012), and we continue with the aim of bringing together philosophy of science and philosophy of technology.

You will remember from previous chapters that PI is motivated by the "information revolution" (see Chapter 2). Science has repeatedly changed how we understand the world, and also ourselves. Now it is revolutionising our abilities as makers of the world. Let us see why. In a nutshell, the information revolution revitalises ancient questions about the relation between nature to be passively observed and known (which the ancient Greeks called *physis*) and practical science or art that interferes with the world (*techne*, from which we derive "technology").

To understand this, we need to take a step back and see what vision of the world is presupposed in ancient and in modern science. What follows is not a detailed historical reconstruction but is nonetheless useful to get a feel for the size of the change PI can bring.

Philosophy and science start with the Greeks, their way of looking at the world and of drawing conclusions from their observations. For the Greeks the natural world is known by passive observation. For the Aristotelian scientist, experimentation was not a means to acquire knowledge but just a means to illustrate knowledge already acquired (for a discussion, see Harris (2005, Chapter 1)). The scientist, according to Aristotle, aims to establish "first principles" – science is knowledge of the *physis* through its contemplation. So science and *techne* are different for Aristotle, because science is not practically oriented. *Poiesis*, or making, concerns only the arts, the *techne*. This is in sharp contrast to the modern philosophical conception of science and of scientific method, most of all in the modern use of new tools to acquire knowledge, particularly during experimentation.

Let us now make a very long jump forward in time. Since the Scientific Revolution (c. 1550-1700), the natural world is a world that the scientist actively interacts with and manipulates in order both to know and create. The shift is from an "organic" view of the cosmos, typical of the Greeks and perpetuated in the Middle Ages, to a

"mechanical philosophy" which pioneering scholars such as Francis Bacon, René Descartes, Galileo Galilei and Isaac Newton started to develop. The change has been so profound that "science" no longer connotes merely "knowledge" and "understanding", but now also embodies practical skills. It was with Bacon that science became a *scientia operativa* (Klein, 2008, 2009): this means that to come to know about the world the scientist does not just passively observe the world, but also interacts with it. The modern scientist is a maker; among other things, she performs experiments, she actively manipulates factors to find out what causes what (Ducheyne, 2005). Experiments, in Bacon's view, are tools to acquire new information, but they are also tools to test theories, according to Galileo.

Performing experiments is thus a way to make, to build, or to construct knowledge. This requires carefully controlling the study, and often requires building specialised equipment, including whole environments such as at CERN. You can easily see how much this differs from an ancient understanding of *physis* (nature), as being discovered simply through passive observation of the world.

We can now summarise the two major innovations introduced by scholars of the Scientific Revolution as follows:

(i) in order to know we need to make; and
(ii) what we know is going to be of some practical use.

These are, in short, the strongholds of the concept of technoscience. A large part of current science is indeed technoscience. Scientists at CERN can run their experiments only thanks to the very sophisticated technology of the collider, and the extraordinary data-processing functionality of supercomputers. Molecular epidemiologists and biologists can study biomarkers thanks to complicated machines that analyse samples of blood, urine etc. and detect immensely small entities. Astrophysicists can study Martian rocks only thanks to the machines that can travel in space and collect samples (those that don't crash). So it is difficult to distinguish fundamental science (knowledge of the basic principles of nature and life) from the technologies that both stem from it and participate in it. We have to include understanding our technological interfaces with nature in our understanding of the practice of science. There is also another aspect to consider. Neither science nor technology, alone, makes any progress. Behind science and technology there always is a scientist, a technologist, or indeed a technoscientist. So let us try to understand what kind of agent is the technoscientist.

The homo poieticus. Before we reveal who is the technoscientist and what s/he does, we need to introduce a more general concept: the *homo poieticus*. We see elsewhere in the book how PI offers an alternative framework to traditional ethical theories (see Chapters 4 and 5). We add here a discussion of what agent the ethical agent is, in the era of technology. The *homo poieticus* is the ethical agent in the era of technology: s/he is the creator of the situations subject to ethical appreciation. This already delineates a constructionist framework: the ethical agent, i.e. the *homo poieticus*, constructs the situations s/he is in. For example, in constructing the internet, we create a whole system. This step goes beyond traditional ethics, where we discuss what is right or wrong or whether the consequences of an action are acceptable or not, but do not discuss who is the ethical agent and how s/he got there. This should make clear why a "constructionist framework" is better suited to the new environments created by technology: technology creates environments and the *homo poieticus* is their creator. Now we know who the agent is. The advantage of a constructionist ethics lies in the fact that, unlike traditional ethics, it does take into account the various circumstances that led the agent to be in the situation s/he is facing. Instead, traditional ethical accounts, whether in the framework of consequentialism or virtue ethics,

take the situation as "given", so to speak. The philosophy of information redresses this aspect: traditional ethical frameworks neglect what is perhaps the most important feature of the ethical agent in the information era, namely her poietic skills. In creating a new system like the internet, *homo poieticus* can make many more choices about who the system favours – and who it may disadvantage. These need to be addressed. Here comes the link with science and technology: the technoscientist is a *homo poieticus* too.

Homo poieticus as technoscientist. There are two things that the technoscientist creates: crafts and knowledge. Let us examine these in turn.

The creation of crafts. The technoscientist produces the "objects of technology" e.g. computers, nuclear weapons, medical devices. In general, these are humanly fabricated artefacts. Traditionally, Lewis Mumford proposed a categorisation of technological objects that included utensils, apparatus, utilities, tools and machines (see for instance (Mumford, 1934)). Later on Mitcham (1994) added to Mumford's categorisation the following: clothes, structures, and automata or automated machines. This list of technological artefacts includes "tools of doing" and "tools of making" alike. Needless to say, there are interesting remarks to make about the distinctions between "tools of doing" and "tools of making". Also, a lot can be said about alternative categorisations of technological tools. The phenomenology of artefacts is invaluable in providing tools to investigate, for instance, their personal or societal effects, or the way artefacts may extend human capabilities and, consequently, alter our experience with the external world (Ihde, 1979). But this is not our concern at the moment. What interests us the most is that technological objects – crafts – are the products of the poietic activity of the technoscientist. In other words, the technoscientist is essentially a *homo poieticus*. The philosophy of information initially conceived of the *homo poieticus* as the creator of e-nvironments, but we can now extend the notion to the technoscientist because he creates too.

The creation of knowledge. There is another aspect of the poietic activity of the technoscientist that is of relevance here: the technoscientist creates knowledge. The insights about the technoscientist and his poeitic activity in constructing knowledge can be found in the kind of epistemology that is part of the philosophy of information. The philosophy of information investigates the relations between the natural world and information (Floridi, 2011c). Such relations will be specified within the network of information, what we might call "constructionist epistemology". Let us step back.

Recall that the information revolution is about our coming to see that we are informational organisms, or *infogs* in the information sphere or *infosphere*. This means that information is key in understanding ourselves, the world, and ourselves-in-relation-with-the-world. Consider now the relation between information and the natural world. The question ultimately concerns the localisation of information: whether there can be information without informee, and whether information can be naturalised in the sense of the semanticisation of data (see Chapter 3). This is a concern for epistemology, and not a new one. There is a sense in which Kant, the German idealists, and the British empiricists were trying to do just that: to understand how we know what we claim we know about ourselves and about the external world (if there is one).

So the technoscientist creates knowledge because it is through our conceptual and experimental tools (in short, our scientific practices) that we make up and systematize our observations (passive and active) about ourselves, the world, and ourselves-in-relation-with-the-world.

The technoscientist embodies all these aspects of poiesis at once: the creation of crafts, of knowledge, and also of the situations we are in and that are subject to ethical evaluation. After all, many results of scientific discoveries have an important ethical dimension. We have the knowledge and the technical ability to build nuclear and chemical weapons – but should we? We have the knowledge and the ability to help women over 60 years of age get pregnant – should we? We can test drugs and cosmetics on animals – should we? And so on.

9.6 Conclusion

We have looked at how PI can illuminate existing debates in philosophy of science, including modelling, evidence and expert knowledge, before moving on to show how the understanding of the *homo poieticus* that PI yields extends interesting issues to address regarding science. You might be one of the many scientists interested in the conceptual design done by the sciences. But if you are a philosophy student, this does not mean that science has nothing to do with you. First, you will encounter science more often than you think. It is often the source of the conceptual change that calls for philosophers as conceptual designers, and there is also a lot philosophy can do with science: building concepts and methods together. The philosophy of information makes a contribution to this when it rethinks subjects such as model validation, evidence, or expert knowledge. One last thought about *poiesis*. The technoscientist creates knowledge, we said. But the philosopher is not very different, in a sense, because the philosopher creates concepts. In a lot of this book we have done conceptual design, which is philosophy. It is this poietic activity that unites ethical agents, technoscientific agents, philosophers, and in the end unites all of us: the poietic dimension inheres in every aspect of our lives.

9.7 Exercises

Find one or two articles in the news (e.g. bbc.co.uk/science), read them, and answer the following questions:

1. What kind of language does the article use to describe research findings? Is it in terms of cause-effect relationships? Is it in terms of hypotheses?

2. Can you retrieve information about the *methods* used from the article?

3. If ethical issues are raised, is there any information about alternative methods to carry out the same research?

4. Does the article mention the debate in which the research is included? Can you figure out whether there is support, dissent, or interests at stake from the scientific community?

5. What do you think is the greatest achievement of science? Is it figuring out the movement of the planets, putting a man on the moon, or putting a smartphone – i.e. a mini computer – in every hand?

9.8 Suggestions for the exercises

1. Look for words like "hypothesis" and "cause" of course, but also look for closely related words like "produces" or "increases" and so on. Watch out for the word 'link'. It is usefully vague, which makes it easy for the media to use. Every time you see it, ask yourself what it means, and what the evidence given for it could support.

2. For instance, does the article mention whether Randomised Controlled Trials or some kind of observational studies were conducted?

3. For example, you might see an explanation that using a Randomised Controlled Trial was impossible, and this is why an observational study was conducted.

4. Look for the names of research groups and funding bodies. Do they agree or disagree?

5. Figuring out the movement of the planets is observational, discovering the world as it is. Putting a man on the moon is partly an attempt to discover nature as it is, by reaching parts of it we could not reach previously, but is also partly poietic, because we had to build the technology to make the journey. Putting a miniature computer in every hand is entirely poietic. Which do you think is most important, and why?

9.9 Further reading

Olsen, Pedersen, and Hendricks (2009), Russo (2012).

Chapter 10

10. COGNITION

Information processing and thinking

10.1 Introduction: What is cognition?

'A human conversation depends on many processes which a scientist would call 'mechanical', in the sense that only physical categories of cause and effect are needed to describe and explain them. … Now, until the chain of explanation reaches the nervous system, nobody minds its mechanistic flavour. True, it has made no reference to the meaning of what is being said; but this, we might say, would obviously be premature. Questions of meaning need not arise until we bring in the human links in the chain. … It looks as if the meaning of a message can be defined very simply as its selective function on the range of the recipient's states of conditional readiness for goal-directed activity; so that the meaning to you is its selective function on the range of your states of conditional readiness. Defined in this way, meaning is clearly a relationship between message and recipient rather than a unique property of the message alone.' (MacKay, 1969, pp. 20-24)

What is cognition? This is a difficult question to answer precisely. It is a question similar to *what is life?* or *what is society?* Answers to such questions have been proposed, but they barely affect how biology or sociology is done. For many scientific goals an intuitive, imprecise idea is sufficient. The same is true for cognition. Cognitive science does not need an absolutely precise answer. It can get by with prototypical examples, and several basic ideas for most problems. We will do the same here. Let us look at some of the simpler organisms that cognitive scientists investigate.

Meet Portia from Australia (*portia fimbriata*), a kind of jumping spider, who is hunting another web-laying spider that we'll call Orba. Most jumping spiders have amazingly good eyesight, unlike other web-laying spiders who rely mostly on senses of touch. Portia uses a very unusual and sophisticated tactic to hunt Orba. You see, Orba's web is not only its (temporary) home and catch net; it is an extension of her sense organs. Orba can locate and identify various objects on the web based on the nature of the vibrations and the way the vibrations propagate along the strings of the web. Portia must climb the web of Orba, and thus she is bound to be detected. What is worse, Orba wouldn't mind snacking on Portia herself. Here comes the strategy: Portia initiates vibrations on the web in an attempt to control Orba's behaviour. Portia, of course, hunts many types of spiders, so she does not know what signals Orba can discriminate between and what behaviour they induce. So, Portia initiates a random sequence of vibrations and observes Orba. Once a signal has attracted Orba's attention, Portia starts repeating it.

Figure 7: Jumping spider

If the behaviour is not conducive to a successful hunt (or may lead to the reversal of the direction of the hunt), the random vibrations are initiated again, until a new behaviour-modifying signal is found.[29] The goal is to make Orba come sufficiently close so that Portia can jump and stab the prey with her venomous fangs – yummy indeed!

Scientists studying insect cognition get very excited by such behaviour because it hints at many cognitive skills we "higher" creatures possess – learning, representation, attention, memory, strategic pursuit of remote goals. It is, of course, too easy to speculate about the "mental life" of such creatures, and the use of such concepts may or may not be fruitful. Let us agree on this much: such creatures possess the rudimentary blocks of cognition and are of interest to cognitive scientists. They belong to the realm of cognition. We want to focus on another aspect of the story: the role of information.

Let's look at Orba first, because her skills are also quite remarkable. Orba's small brain (400K – 600K neurons) can discriminate a significant set of distinct patterns of vibrations. The discriminating networks of neurons cause other networks to initiate distinct behavioural patterns. Some of these patterns are further modulated by other sense organs or by the internal metabolic state of the spider. There may be a pheromone discriminated by the spider's olfactory system, or the spider may be hungry. Some behavioural patterns change the relation between the spider and the sources of the vibrations. This may help the spider position the source and initiate an attack or mating behaviour (or something else). The spider's nervous system, together with the behaviour it generates, generates a *level of abstraction* (see Chapter 2) that structures its environment, and allows it to identify remote objects and let their properties influence its behaviour.

Portia pushes such complex relations between the environment and the organism, mediated by the complex structure of neural connections in its brain, a step further. The brain, when recognizing the context of the hunt

[29] The behaviour of spiders from the genera Portia is actually more sophisticated. For example, they may approach the prey faster if wind disturbs the web, causing too much noise and blocking the vibration signal. Or, they may abandon the attack via the web, and using some of the observations from the web, initiate an aerial attack, dropping from a nearby branch at the optimal angle to avoid detection. See Wilcox and Jackson (2002) for more discussions.

for an unfamiliar spider, with the discriminatory capacities of the visual pattern recognition system, initiates behaviour (the random vibrations) whose consequences for Orba are monitored through the visual system, and adjusted (by switching to a repeated vibration pattern) when Orba's neural control mechanisms are hacked. In the process, Portia generates a level of abstraction not only to the properties of Orba, but to the properties of Orba's behaviour. Portia discovers a relationship between vibrations and the behaviour of the type of spider Orba is (according to the level of abstraction defined by the prey recognition mechanisms). The discovered relationship is "recorded" in new neural connections that link patterns in the environment detected by its sensors to behaviour. This is the result of the complex mechanism in the spider's brain that "processes and integrates" the input patterns into new sets of distinctions and control relationships between senses and behaviour.

We were careful not to use informational language in the description of the situation; however, we see that the important elements of an informational story are staring us in the face. The story suggests that thinking of the spiders as organisms extracting, processing and using information (but not only information) in their interactions with the environment allows us to understand their behaviour more fully. Indeed, the sensory system and neural pattern recognizers structure the environment into data sets based on the levels of abstraction. The feedback mechanisms use external objects as sources of information to which the data sets apply. When Orba senses vibrations as belonging to a stable datum coming from a given direction, and when Orba's own movement reveals that the source of the vibrations is independent of her, then she can be described as extracting information from an external source and using that information to control her behaviour. When Portia relates the information about the behaviour of Orba to the information about the vibrations she produces, she obtains information about how Orba can be controlled. In this case, Portia can be said to convert one kind of information (with one set of data based on one level of abstraction) into a different kind of information (with a different set of data, based on a different level of abstraction). Portia is processing information and extracting latent information about the environment from limited information inputs.

In this story, the informational description has semantic elements – elements concerning the meaning of information. In current debates in the philosophy of information, the question about the exact conditions leading to semantic information is not settled. It is not agreed upon whether the spiders would count as using semantic information. We will return briefly to this problem in the last section of this chapter. The nature of the meaning of information was discussed in more depth in Chapter 8. For now, it is sufficient to recognize that this level of cognitive organization is a contender for semantic information, even if the use of semantic language is only metaphorical. The fact that it is useful and revealing, even if metaphorical, suggests that cognitive systems are the right home for information.

This is the attitude of the majority of cognitive scientists. Informational language is freely used in describing cognition in general and the operations of specific cognitive mechanisms. Little attention is paid to the precise conditions of use and the appropriateness of informational locutions. As philosophers, this should bother us. However, we should be careful not to throw the baby out with the bathwater. We should not end up at a place that denies the concept of information to cognitive science. If philosophy of information sterilizes the concept so much as to make it applicable only to a limited domain of human behaviour, it is more likely that the sterile philosophical theories will be made irrelevant, than that the language of cognitive science would change. The concept of information is simply too useful.

So what is cognition? Let us leave it to cognitive scientists for the moment and offer a few attempts to define it:

> Cognition refers to the mechanisms by which animals acquire, process, store, and act on information from the environment. These include perception, learning, memory, and decision-making. (Shettleworth, 1998)

> [A]ll organisms, including bacteria, the most primitive (fundamental) ones, must be able to sense the environment and perform internal information processing for thriving on latent information embedded in the complexity of their environment.... We then propose that by acting together, bacteria can perform this most elementary cognitive function more efficiently as can be illustrated by their cooperative behavior (colonial or inter-cellular self-organization).
> (Ben-Jacob, Shapira, & Tauber, 2005)

> Here the term cognitive refers to processes of acquiring and organizing sensory inputs so that they can serve as guides to successful action. The cognitive approach emphasizes the role of information gathering in regulating cellular function.
> (J. A. Shapiro, 2007)

Such definitions are too vague, but they clearly demonstrate the importance of the concept of information. The relationship between information and cognition is still unclear (we will examine it in more detail in the following sections), but it is evident and undeniable. Cognition is the home of semantic information – we have the address. Now we need to open the door!

10.2 The birth of cognitive science

The birth of cognitive science was the result of an unhappy marriage between the Freudian and the Turing revolutions (see Chapter 2 and a different treatment of the importance of Turing in Chapter 13). Further, it was baptized by the early development of the mathematical theory of communication and cybernetics. As such, it was the first genuine natural science of the informational turn. By the 1930s, the revolutionary idea of Sigmund Freud, that most of the operation of the mind is the result of a hidden layer of subconsciousness, inaccessible to reflection, had run into difficulties. The difficulties were not related to the idea that there is something hidden in the mind, but to the way Freud and his followers in the psychoanalytic tradition had theorized about the subconscious. The problem was that psychoanalysis had grown so speculative and void of empirical content that it began to be regarded as unscientific. The scientific community was disillusioned by trying to understand the working of the mind and had converged on the idea that science has no business venturing into the mental black box. The only scientifically acceptable domain of investigation was human and animal behaviour. This was the birth of "behaviourism", which between the 1930s and the 1950s was the predominant school of psychology. The behavioural approach achieved its most systematic prominence with the work of B. F. Skinner (Skinner, 1938; Skinner & Ferster, 1997), who never accepted cognitive science, claiming that 'cognitive science is the creation science of psychology.' (Skinner, 1990).

It is important to note that early neural science was developing alongside both psychoanalysis and behaviourism. Before the turn of the twentieth century, the work of Camillo Golgi and Santiago Ramón y Cajal led to the careful description of the structure of the neuron and the hypothesis that the neuron is the basic functional unit of the brain. This work resulted in a shared Nobel Prize in Medicine in 1906. Unfortunately, neither Freud's followers nor Skinner's could connect the study of the neuron and the neuronal networks in the brain to psychology. It was not clear how to connect the neuron to the idea of mental or cognitive function. This is the first place where the informational approach became relevant. Already in the late 1930s and 1940s cyberneticists, like Norbert Wiener and Warren McCulloch, attempted to understand the functioning of the

mind on the basis of structural principles of control and information. McCulloch and Pitts (1943) developed the first functional model of artificial neural networks. Some of this work happened alongside the work of Shannon in information theory. But the conceptual marriage between the work of Turing, the work on formal theories of information, and the work on theories of neural science, took place with the work of John von Neumann, arguably one of the great scientific geniuses of the twentieth century. In a set of lectures in 1955 shortly before his death, which became the book *The Computer & the Brain,* von Neumann (1958), suggested that the neurons in the brain may be interpreted as performing digital computation. Thus, Turing's work on the theory of digital computation had direct significance for understanding the mind. The hypothesis was that the brain is a digital computer that is implemented by the brain's neuronal structure. The model is as follows: the brain receives information from the environment through the senses. This can be regarded as input data to the system. The brain takes this input and processes it to produce an output. The idea of digital computation can be used as a framework for investigation of what happens between the inputs and the outputs. This is what was missing from behavioural theories, which focus only on stimulus and response relations – the inputs and outputs, in the new terminology. The brain, then, according to this new model, has a specific function – information processing – which has the effect of connecting stimulus to response, but now the function may be investigated on its own terms, as a computational process.

This union of the theory of computation, ideas from the formal theories of information, and neural science, defined a new research program for investigating what happens in the black box of the mind that is rigorous and scientifically respectable, unlike the psychoanalytic tradition. The program initiated by von Neumann did not turn the tide of psychology on its own. It is important to mention the work of yet another twentieth century genius, Noam Chomsky. In 1959, Chomsky published a review of Skinner's book *Verbal Behavior* (Skinner, 1992), which became better known than the book itself (Chomsky, 1959). In it, he outlined the major problems with behaviourism, and made a compelling case for the importance of studding the internal workings of cognition. This, together with Chomsky's then recently published book *Syntactic Structure* (Chomsky, 1957) which revolutionized linguistics, became the last nail in the coffin of behaviourism, and opened the doors for cognitive science. Chomsky's influence became more important for linguistics and empirical cognitive psychology, while Turing and von Neumann's work became central for the emergence of the field of computational cognitive science and artificial intelligence. Chomsky's work did not contribute directly to theories involving the idea of information, so it will not be discussed further here.

The new science of cognition, fuelled by a renewed courage to venture inside the mind (and a good deal of enthusiasm), branched in two directions. One direction was the new link between psychology and neuroscience (which until that point was viewed as a branch of physiology). The simultaneous development of computer science led to the development of computational neuroscience. A second direction was the program of *artificial intelligence* (AI). Both branches made use of notions of information. Here, however, we shall focus on the foundational problems related to AI as they link more directly to the problem of the philosophy of information.

10.3 Computationalism

The program of AI has always had two goals: (1) to produce a machine that may exhibit general or human-like intelligence, and in the process (2) to reveal something general and fundamental about intelligence and cognition. The first goal is mostly related to engineering. Of course, it is deeply connected to philosophical problems related to what intelligence is, how it can be recognized (remember the discussion of Turing's test in Chapter 2), and whether AI is achievable in principle. Here we focus on the second goal.

The von Neumann model quickly became entrenched in the work in AI. Target AI systems were organized as follows: the system is given an input state of the environment (e.g. the position of pieces on a chess board), the system performs operations on the input state (e.g. explores different combinations of moves and evaluates the new possible states of the board), and selects a desired output operation (e.g. a next move). In other words, the von Neumann architecture for computation follows the pattern: *input→processing→output*. This model was translated into a general model of cognition: the cognitive system receives an input from the environment; it then processes the input, and selects an action based on the input, current state, and its history. This is described as the "horizontal architecture" of cognition (Fodor, 1983), and follows the general three-stage sequential pattern: *perceive→think→act*. The cognitive system takes information from the environment (i.e. perceiving), the information is processed (i.e. thinking), and an action is produced. For this reason, this model is also described as the "information processing" model of cognition, as the most important stage of the model, *thinking*, is associated with the *processing* of the information.

As a matter of fact, the information processing, horizontal model is more general than early AI assumed. We shall see alternative versions below. Early AI adopted a more specific principle, described as the "physical symbol systems hypothesis" (Newell & Simon, 1972, 1976).[30] The input to the system is given as a collection of symbolic expressions depicting the state of the environment, and the system transforms the expressions by following *rules for transformation*. Then, using methods of *heuristic search*[31], it explores the space of possible representations of the environment with symbolic expressions to determine the best outcomes of its actions compatible with the input expressions. This guides the system in selecting the next action. Consider the problem of chess again. In a chess-playing system the environment is the state of the chess board. It is represented by some data structure (say, an 8x8 matrix of figure positions – a kind of symbolic expression). The goal is to generate a move that will lead eventually to a checkmate. The system would search the space of possible sequences of moves that will lead to a win. Such a space, however, is hyper-astronomically large. The system must by necessity find ways to reduce its explorations dramatically. It will explore only a limited number of possible game developments: it will use rules of thumb, databases of known games, even information about opponents. These are all search heuristics. In the process, the system will use many axillary data structures (symbolic expressions) that do not represent the environment. The difference between the chess-playing app on your cell-phone and Deep Blue, the system that beat the world champion Garry Kasparov, is in the heuristics and axillary data structures.

To explain further, in the current language of philosophy of information, the hypothesis of early AI was that cognition is analyzable at a specific level of abstraction, distinct from its physical implementation. This idea proved a methodological maxim for investigating the operation of actual cognitive systems, outlined most succinctly by the computational neuroscientist David Marr (1954-1980). In his posthumously published, highly influential book *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information* (Marr, 1982), Marr recommends that cognitive systems in an organism, such as the visual system, should be investigated at three levels: "*function*", "*algorithm*", and "*implementation*". The level of function is a level of abstract description of what task the system has to perform. Thus, the chess-playing system has to identify an optimal

[30] The "physical" in the physical symbol system hypothesis is the idea that intelligent, cognitive systems, as physical systems situated in the world, are intelligent because they implement (as hardware) the operation of an abstract symbol system (as software).

[31] The space of possible representations of the environment can be astronomically huge. For example, if the environment can be described by 400 binary parameters (a one mega-pixel image from a cheap digital camera is described by 3,000,000 binary parameters), then there will be 2^{400} possible representations of the environment. This is more than number of elementary particles in the visible universe (primarily photons and neutrinos). It is impossible in practice for a physical symbol system to explore all representations. The search needs to be guided by *heuristics* (rules of thumb, roughly) that only explore a limited set of possibilities. The idea of heuristic search was central for the practicality of physical symbol system hypothesis as proposed by Newell and Simon.

(or very good) move from the current state of the board. The level of the algorithm is a description of the specific procedures and representations (i.e. expressions) that the system uses to compute the function. At this level one specifies how the chess board is represented, and what steps the system takes to discover the optimal move. When one programs a chess-playing system, this is the level at which one works. The level of implementation is a description of how the physical system implements the steps of the algorithm. This could be a description of how a computer is built, or (as it is interesting for cognitive systems) it could be a description of how the neurons in the brain decode and perform the algorithms. Such three-level analysis is natural for human-designed software systems, but Marr's claim is that it is necessary for all cognitive systems, including biological ones. Our goal now is to see how the idea of information enters the picture. Later, we shall look at objections to this program of AI, and we should be able to evaluate whether informational ideas transfer to the suggested alternative approaches to cognition (and AI).

The idea of information enters at two places. One is related to the input, the other to the processing. It is assumed that the input to perception is not just anything, but some information about the environment. Consider the pattern of photoreceptors on the retina of the eye. Photons are reflected from an object in the environment; as a congruent stream they enter the eye and are focused on the retina. The variation of incoming photons is imprinted on the patterns of photoreceptors, in such a way that structure from the object is preserved. We can think of the photoreceptors as cells in a linear array, and their variable excitation levels as the values of the cells. If we think of the possible excitation levels as being of discrete values, the array can be viewed as a digital array of bits. This array of bits, then, is correlated with the objects in the environment. The upshot is that if a vase is in the visual field of the eye, one pattern of bits is present. If a dog is in the visual field, a different pattern is present. Following this analogy, it is possible to think of the relationship between the retina and the objects in the environment in terms of Shannon's model of communication (see Chapters 2 and 3). The environment is a *source* of information, the retina is the *receiver* of information, and the intermediate medium (the photons) is the *messages*.[32] (As we shall see below, this model of perceptual input is criticized by some, such as Gibson, but in early AI it was considered the natural model of cognitive input.)

The cognitive system, physically, is set up in such a way that its perceptual system enters into an appropriate "informational relation" with external objects by taking advantage of natural physical regularities (e.g. the physics and geometry of light). In virtual artificial systems, such regularities are assumed in the background, while one works directly with inputs. The model assumes that any semantic role of the internal data structures is specified by the "informational relation". If a data structure is regarded as representing a dog, it is because of its connection to input patterns that track dogs in the world. The actual work of the AI researcher and the cognitive scientist, however, should be focused on what happens after: processing.

This is primarily why the model is called the information *processing*, rather than the "information *acquisition*", model of cognition.

Now let us turn to the question of processing. Remember that von Neumann's suggestion, connecting Turing's theory of computation to neuroscience, was based on the idea that the network of neurons in the brain *implements* a computing device working on the sensory input. Strictly speaking, this is a tall order. To say that a system implements a computing device is to say that the natural dynamics of the system (whatever happens internally and as a result of external influences) is sensitive only to distinctions in the inputs – what we have called *data* – and any other perturbations, somehow, get absorbed or filtered out. It is not that the dynamical

[32] The precise location of the receiver, naturally, can be moved further into the brain. It could be a part of the visual cortex, in which case the retina will be a part of the message.

behaviour of the system – its behaviour over time – is *entirely* controlled by the distinctions in the inputs; many interesting things may happen in the system. But the system keeps the distinctions from the inputs, and anything that happens as a result, separate from other influences. For example, neurons in the brain engage in active metabolism, consuming vast quantities of nutrients and oxygen. Yet, the idea is, these metabolic processes of the neurons (what they do to stay alive) are kept separate from the computation processes (what they do to transmit information).

If we think of the internal working of the system as a computing device, then we can think of the system as a device that processes the input (i.e. data) to generate some output. The idea of "information processing" is really the idea of "data processing", with the added assumption (at least in the case of cognition) that the data are associated with the external environment. Marr's three-level analysis assumes that the story of cognition includes a level (the level of algorithm) that is entirely based on data processing.

There are two things we should notice about the way the idea of information enters the early AI program. First, while both Shannon's communication model (see Chapter 2) and the idea of data processing (called information processing) are present, the actual mathematical theory of information does not play a central role. Second, the observation that cognitive information processing is really data processing makes it more difficult to understand the role of semantic information for cognition. The question is: how can performing processing operations on input data to produce output data involve the idea that the data is meaningful? Where is the meaning coming from? The chess-playing system operates on various abstract data structures, such as matrices, that may be regarded as representing chess pieces on a board by us, the designers. But why should they be regarded as chess pieces by the system? Is there a sense in which their representing chess position makes a difference for the actual information processing? It seems that semantics cannot come from data processing alone. It must come from somewhere else.

A possible place for semantics is the relation between the input and the environment. This is more difficult to see in the case of a chess-playing system, as chess is an abstract game. Consider again a system for computer vision. These were the kinds of systems Marr was studying. The input of such a system is given by a camera that digitises an image with a CCD chip. A horse is grazing in a field; light reflected from the house enters the camera, exciting electrons on the chip and little zeros and ones are produced. It could be argued that the meaning of the data inside the system comes from the causal process connecting the horse to the input. (We looked at causal theories of meaning more closely in Chapter 6.) Can this suffice as an account of semantics for the information processing model? Not really! The problem is that the internal operations of the system depend only on the state of the input. The causal origin of the input is actually irrelevant for the operation of the system. The system does not care that the input was generated by a horse. This is because there is no sense in which the connection between the horse and the image can influence the operation of the system independent of the input. The problem is not of causality alone. It springs from the entrapment of the information processing system between its inputs and outputs.

10.4 The connectionist alternative

In the 1980s there (re-)emerged a different architecture for the operation of cognition: the "connectionist architecture". This attempted to stay closer to the physiological structure of the brain as a network of neurons. It was based on the idea of an "artificial neural network". As we noted earlier, theoretical investigations of biologically-inspired network systems date back to the 1930s, coming from early work in cybernetics. The

1980s, however, brought new techniques for constructing such neural networks that can accomplish complex tasks.

The basic idea of a neural network (as used in AI research) is a network of elements called *neurons*, which are connected to other neurons by *links*. Each neuron may have a bunch of *input links* from other neurons, and a bunch of *output links* to other neurons. Neurons that have no input links are called *input neurons*, and neurons that have no output links are called *output neurons*. Each neuron can be excited to a given level, and how excited a neuron is depends on the input links it receives. We think of the links as channelling excitation from one neuron to another, and each link has a *weight* that determines how much excitation it propagates to the next neuron.

Typical neural networks are organized in layers (see Figure 8). On one side is the input layer; the middle contains the hidden layer(s); and the other side is the output layer. The network operates by setting the input neurons to a pattern of excitation values, and letting the excitations propagate to the output. The network, thus, connects patterns of input to patterns of output. Such a network can be used for all sorts of tasks. For example, the network may classify pictures into conceptual categories – the input layer may encode a bitmap, and the output layer may encode concepts, such as "horse", "house" "tree", etc. The amazing thing is that relatively simple networks can make very subtle pattern recognitions.

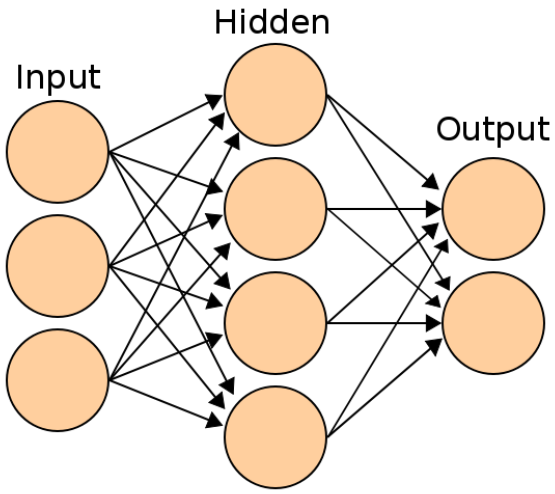

Figure 8: A neural network

In the 1980s, people like James McCelland and David Rumelhart discovered methods for building networks that can accomplish practical tasks. If the links between neurons are fixed, the propagation of excitation may be modulated by specifying the weights of the links. The behaviour of the network can be "programmed" by changing the weights. The methods involve training a neural network on inputs that produce known outputs, and modifying the weight, using special algorithms, so that the network gets better and better at the task. When the network gets quite good at the training set (which may be quite large) it also performs correctly on unknown inputs. It turns out that such networks exhibit many types of learning and performance behaviour seen in biological cognitive systems. Such observations have been used by some to argue that cognitive systems are essentially special kinds of neural networks, and the theory of neural networks should be regarded as the foundational theory for cognition (Churchland, 1992). This strong view is generally not accepted, but it is acknowledged that artificial neural networks are an important research tool, not only for AI, but also for biological cognitive science. In any case, we will not venture deeper in the foundational debate about the scope of connectionist networks. Instead, we should consider how the idea of information connects to this cognitive architecture.

What has changed in the move from early AI to connectionism, and what has remained the same? The general pattern of the horizontal architecture has remained the same. The structure is still *input→processing→output*. The change is in the way we conceive of processing. Thus, the relation to information as input being correlated to the environment is essentially the same. Neural networks are not naturally physical symbol systems (although a symbol system may be implemented on a neural network). They process information in a different way. The

main difference is that they distribute the information along the entire network, or large parts of it, and operate in parallel. For this reason, they are often described as performing "parallel distributed processing" (PDP). In the typical programmed symbol systems, information-carrying data structures are clearly specified. Expressions are local structures, internally organized by rules of syntax. The transformation rules operate on local properties of the expressions. It is usually clear from the outset (at the design level) what stands for what – what carries what information and how the information is encoded. This is exactly what is not readily available in neural networks. Whatever internal structure emerges to act as information carrier in a neural network, emerges in the process of training, and needs to be discovered post factum. In fact, there is an entire field of theoretical neural network research dedicated to the analysis of how information is encoded and processed in PDP systems.

Informational analysis of PDP systems has taken two approaches, one purely syntactic and another with some semantic elements. The first approach is based on applications of Shannon's mathematical theory of information (or more advanced extensions) to the analysis of the operation of neural networks or to the learning algorithms that train the networks. For example, measures such as mutual information can be used to understand whether individual neurons perform similar or different tasks in the information processing. Such measures can be used to adapt network. Similarly, such measures can be used to understand the relation between input and output neurons, with the idea that, in a well-trained network, the information the outputs contain about the inputs is maximized. This approach usually relies on other statistical tools that are used in conjunctions. It is safe to say that, while information theoretic ideas are useful in such analysis, they do not have an indispensable or even central role, beyond the probability theory hidden in them.

The second approach attempts to provide semantic analysis of the operation of the neural networks. Unlike the classable symbolic AI, where the carriers of sematic and representational content are clear, in PDP cognitive architectures the carriers of content are more difficult to identify. The neural network researcher and cognitive scientist Paul Smolensky (1986) argued that in a PDP architecture it is possible to have information processing and representations, only at a "sub-symbolic" level. The sub-symbolic level is identified using abstract mathematical techniques, where the activation state of the network is interpreted as a vector in a multidimensional vector space, and regions in the space are interpreted as the content carriers. Smolensky has developed a sophisticated theory for this analysis called *harmony theory*. The theory has clear connections to Shannon's theory of information. Some philosophers, like Jerry Fodor (1997), have criticized this semantic analysis of the PDP architecture because it is not clear what causal role such "semantic carriers" have in the working of the cognitive system, as they exist only in the abstract mathematical spaces. As a result, Fodor thinks, they have no real explanatory function.

Harmony theory, like Shannon's theory of information, is not a genuine semantic theory. It allows us to understand how information is processed, but not where the information comes from. There is no claim that neural networks, as cognitive architectures, operate with information because they may be analysed effectively with the formal techniques of information theory. Semantic information, if present at all in the architecture, still depends on the relation between the input and the environment. In this sense, connectionism is no better than symbolic AI. The theory, whether harmony theory or some alternative, is still only a theory of data processing. There have been alternative approaches to cognition, however, that have taken a different approach to the input problem.

10.5 Embodied cognition

Remember Orba, the spider who was about to become lunch. Remember we said that Orba's web was a part of her sensory apparatus. Can we realistically expect to understand the operation of Orba's cognitive system if we surround the brain with an imaginary box, call everything outside "the environment", call everything inside "the cognitive system", and call all transitional processes "inputs" and "outputs"? There has been a growing movement of cognitive scientists, philosophers and AI researchers who think that such a box would be arbitrary and scientifically unfruitful. They believe that, except for specially designed cases, like a chess-playing program, the cognitive system is closely dynamically connected to the environment and the physical characteristics of the body of the organism. Any scientific explanation of the working of the cognitive system must include an essential contribution of the body and the dynamical interaction of the body with the environment. Some stronger versions of this approach literally claim the environment is a part of the cognitive system (Brooks, 1991; Chemero, 2009; Thelen & Smith, 1994). What this boils down to is a rejection of the horizontal architecture of cognition. This is replaced by an *embedded architecture*. In the embedded architecture, cognitive processes are not trapped between the screens of inputs and outputs. They enter into more complex relationships with the world.

The earliest more systematic attempt to understand the mind as an embedded system was developed in the Phenomenological tradition between the 1920s and the 1960s, by philosophers like Martin Heidegger, Maurice Merleau-Ponty, and Hans Jonas. These philosophers did not have a direct influence on the development of cognitive science, but in recent years there has been a renewed interest in connecting foundational problems about cognition to ideas from the Phenomenological tradition. At the moment, such investigations make little use of the ideas of information, so we will not discuss them further.

10.5a Ecological approach

The earliest attempt to develop a theory of embodied cognition (although the term was not used at the time) influential in cognitive science is the so-called ecological approach to psychology developed by James J. Gibson (1986). Gibson rejected the horizontal architecture and the "information processing" model. In his theory of perception, he insisted that the bi-directional relation between the environment and the organism is central for understanding how organisms perceive their worlds and how perception controls action. The idea of information was absolutely central for Gibson's theory of perception; much more so than it was for the early AI model. As we saw above, in the early AI model, information enters the system only through the input. The natural counterpart in psychology is the theory of senses. In early post-behaviouristic psychology, perception was viewed as the process where the world acts on the senses (sets the inputs) and then the brain analyses the inputs to generate an internal representation of the environment (it processes the information). This was the basis of Marr's account of perception. Gibson rejected this view. He argued that perception is an active process, where the organism extracts and acquires information from the environment by tracking invariances and systematic changes in the *sensory array* – the stream of excitations a sense organ receives from the environment. In perception the organism constantly maintains dynamic contact with the environment and latches onto and utilizes the invariances directly. The acquisition of information, and thus perception, is not something that passively happens to the organism – it is not an input – but it is something that the organism does, actively, through work – it is a kind of output. We can therefore call Gibson's model the information acquisition model of cognition.

Let us look at Gibson's conception of information more closely. First, Gibson rejects the Shannon communication model (which is not the same as Shannon's formal theory of information discussed in Chapter 2) as appropriate for understanding how information comes from the environment to the organism. According to the Shannon communication model, a source sends a message to a receiver, and then the receiver tries to recover the (probability distribution for the) state of the source based on the message and history of other messages. But the environment doesn't send us messages. There is no sense in which the cognitive system receives something from the environment and then focuses only on the received object. The point of the message, after all, is to encapsulate the information, so that you can put it in your "pocket" and go. Second, in Gibson's conception of information, the source of the information is not independent of the agent. Because the invariances on the sensory array are directly linked to the actions and uses the agent has of its environment, the agent perceives the objects directly as affording certain actions. Space is not just a geometric void; it is something that affords unrestricted movement. Space is different for a land animal and for an aerial animal. A branch is perceived as graspable for a primate, "landable" for a bird, and edible for an elephant. Gibson coined the term *affordance* to capture this idea. Perception, thus, gives us not objects but affordances. Affordances are organism-dependent. Thus, the sources of information for organisms – what may be regarded as data in the world and the levels of abstraction at which the data is analysed – depend on the organism and its interaction with the environment.

Where is information for Gibson? This question does not have an easy answer. On the one hand, Gibson talks about information being in the environment. The light coming to the eye contains information about the world. On the other hand, information is only present in the sensory array when the organism extracts it dynamically from its interaction with the environment. The best way to understand Gibson consistently is to interpret the first claim as about *potential information*, and the second claim about *actual information*. The environment has certain dynamically changing *distinctions*, which can propagate to dynamically changing distinctions in the organism by various media, such as light. The distinctions are then channelled to the behaviour of the organism – they make a *difference* – in a way that the behaviour modulates the sensory array and allows the organism to identify more precise distinctions. Only some distinctions are, of course, relevant for the organism; only those are potential information. The potential information gets realized when it is integrated into the organism's behaviour. Actual information requires an organism while potential information does not.

Gibson seems to be holding (implicitly) a view of information reminiscent of a pragmatic theory of information. Pragmatic theories have been formulated by (MacKay, 1969; Nauta, 1970; Vakarelov, 2010). Pragmatic theories of information hold that semantic information requires an agent interacting with an environment in a goal-directed way, and that information modulates the agent's behaviour. The semantic aspect of information depends on the role information states play in goal-directed behaviour and on the way they connect to the external environment. In contrast, in the model assumed by early AI, only the connection between the states of the environment and the input is relevant for semantics. In a pragmatic theory the connection is important, but it is not the only thing. Accounts of semantic information that rely only on a connection to the environment have one major problem: too many phenomena in the environment are connected. For example, if an input to a system is correlated to air temperature, the input is also correlated to air pressure. This is because temperature and pressure are correlated. Thus it is not clear, looking only at the correlations, what is the correct connection: is it to temperature, or is it to pressure? In a pragmatic theory, the internal role may discriminate between the two connections because, either, pressure and temperature affect the organism in different ways and thus involve different

behaviour, or the behaviour may be insensitive to the distinction and thus the level of abstraction of the information state may not distinguish between pressure and temperature. The aim here is not to go into any depth discussing the different theories of meaning for relational and pragmatic theories of semantic information. Instead, it is to observe that in a conception of information appropriate for an embodied architecture, different opportunities exist for understanding the natural place of semantic information. The syntactic level of analysis of the algorithmic level forces all semantic distinction to be somehow channelled through syntactic relations. The embodied architecture can take advantage of the dynamical relation of the organism and the environment to offer opportunities for an alternative conception of semantic information.

10.5b Dynamical approach

Some proponents of embodied architecture insist that we need an alternative descriptive framework for cognition that moves away from the idea of information processing and towards *dynamical systems theory* (Beer, 2000; Gelder, 1998; Thelen & Smith, 1994). Dynamical systems theory is a mathematical framework commonly used in physics. It describes a system in terms of a collection of variables (again, a kind of level of abstraction) and differential equations among these variables. The possible values of the variables form a *state space* for the system and the change of the behaviour of the system in time is represented by *trajectories* in the state space (see 9). The differential equations, essentially, determine the trajectories. Dynamical systems theory offers new

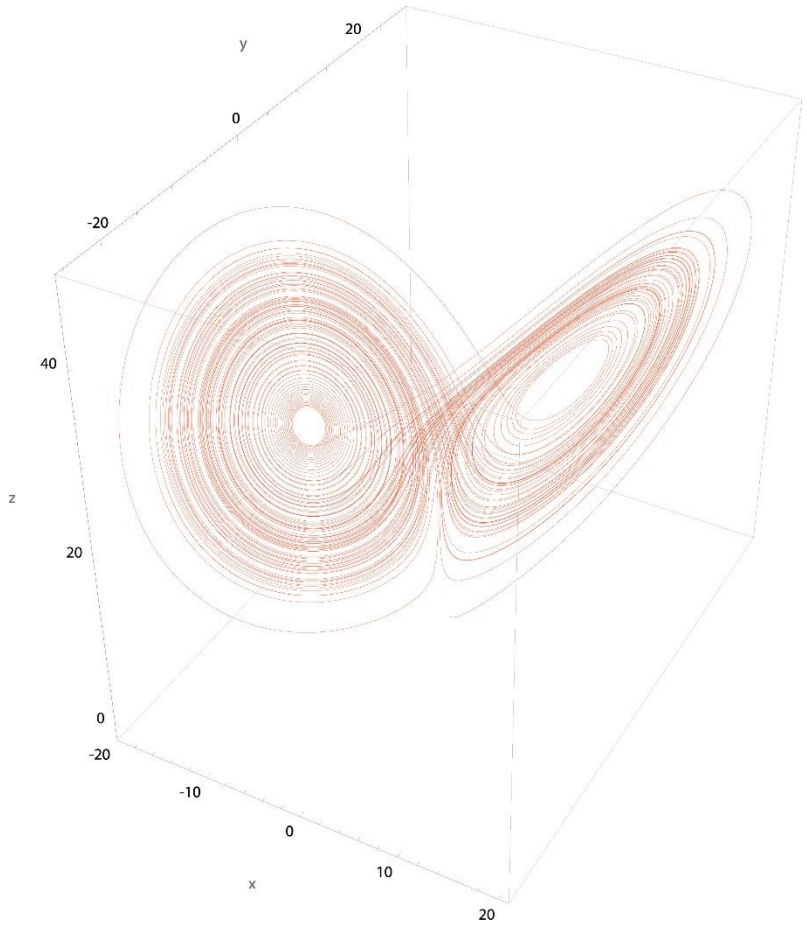

Figure 9: The famous Lorenz attractor is an orbit in a state space.

methods for describing dynamical phenomena that are based on how the system flows through its state space, and on how the flow changes with the change of the initial condition of the system. For example, some collection of states may act as *attractors* in their neighbourhood. When the trajectory of the system enters near the attractor, it tends towards the attractor no matter where it came from. An attractor, therefore, may be used to explain why the system exhibits some stable behaviour. Figure 9 shows the famous Lorenz attractor. Sometimes, the flow of trajectories about a point may diverge in two different directions, a bit like a river splitting in two. Very small differences in the initial conditions of the trajectory may result in big differences in behaviour. There are many other related concepts definable in a dynamical system that are difficult to define otherwise (other examples include *chaos, periodic orbit, strange attractor,* etc.)

There are two places where the use of dynamical systems theory in describing cognition differs from the other approaches, including the symbolic and connectionist, but also the ecological: first, of course, the use of new methods; but also, second, the liberal choice of variables – the variables used to describe a system may be arbitrary. For example, a variable may be the average mass of the system, the distance between two objects, the period of a repeated movement, as well as more typical variables, such as the position of an object, the excitation level of a neuron, or the frequency of light at a photoreceptor. A dynamical system theoretic description usually looks for the smallest number of variables needed to describe the interesting aspects of some behaviour (without prejudice about what these variables are), and regards the discovery of such a small set as explanatorily important. In many interesting cases of cognitive behaviour, the variable may track complex aggregate properties of systems that include parts of the body and environment. These observations have been especially important for proponents of the embodied program. The fact that the effective description of the cognitive system involves variables that amalgamate the brain, body and environment, suggest that cognition cannot be shoved in a brain, while the world is reduced only to inputs to the brain and a space for actions. It is argued that cognitive behaviour cannot be effectively understood by looking at the brain, body and environment as distinct units connected by input/output relations. Instead, cognition emerges as the joint interaction, a joint dynamical dance, of the different components.

Some proponents of the dynamical approach to cognition aim explicitly to eliminate concepts such as information or representation from the study of cognition. A movement, describing itself as *radical embodied cognition*, holds that all cognitive phenomena, including high-level thought and language, may and should be investigated only with dynamical systems theory (or similar non-informational techniques (Chemero, 2009; Gelder, 1998). A more moderate movement insists that many lower-level cognitive tasks may be completed without information and representation, but some *representation-hungry problems* require representation and symbolic processing (Clark, 1998). Examples of such problems include: high-level abstract reasoning, reasoning about non-existent objects, or about alternative states of affairs (counterfactual reasoning).

Much of the debate between the different movements of embodied cognition centres on the idea of representation. However, the idea of information is carried along. We will not go further into the question about whether and how representations enter into cognition. The question of representation will be explored in more detail in the next chapter. The question about information is important, however. One problem with this debate is that the notion of information is too readily connected with the notion of symbol processing used in early AI. Thus, the objection is to "information processing" as used in the version of the horizontal architecture exemplified by the physical symbol system hypothesis. As we saw in the discussion above, the notion of information is more general than, and fundamentally distinct from, symbols. An important alternative is not considered: that a more general notion of information and "information processing" may be needed to investigate the operation of complex dynamical systems, the kind of systems we expect to see in cognitive agents.

In dynamical systems research there is a growing interest in informational analysis of the operation of cognitive systems (Williams & Beer, 2010). Such research spans both artificial intelligence and neuroscience. The research is based on mathematical techniques from more advanced versions of information theory than that developed by Shannon. Such techniques use various information measures to track how information enters the system, how it propagates among different parts of the system, how it is integrated from and split into different informational streams, and how it modulates the system's behaviour. The approach uses information theoretic concepts to define operation principles for such systems: for example, entropy minimization (or maximization), redundancy reduction, or encoding optimization.[33] The idea is to understand how the system learns from the environment and manipulates or transforms its input by regarding the system as performing such informational operations. The lesson should be that the idea of information is not only compatible with the dynamical approach, but is actually essential for understanding the complex dynamical systems inside the brains of cognitive agents. The idea of information promises to bridge the gap between the lower-level analysis of neurons and dynamics and the higher-level world of the mind.

10.6 Conclusion

In this chapter we looked at how the concept of information has been relevant for the investigation of cognitive architectures. We saw that the birth of cognitive science was deeply entangled with the concept of information. The first systematic attempts to investigate cognition abstractly were closely related to the study of AI based on symbolic operations. Such an approach conserved cognition as a computation process and thus there was a natural connection between cognition and information processing. Alternative cognitive architectures not based on symbolic processing were also proposed. We considered connectionism and several types of embodied approaches. We observed that the idea of information plays an important role in such alternative approaches. Attempts to dismiss the importance of information for cognition rely on a narrow conception of information related to symbol processing and digital computation. An important lesson of PI has been that the idea of information is more general than the idea of digital computation. It is this more general and more fundamental idea of information that is most important for the study of cognition. In a sense, information as a fundamental concept becomes more important once cognitive science moves away from the symbolic approach. This is because the symbolic approach only really needs the narrower idea of digital computation.

Another lesson of this chapter was the difficulty of capturing the idea of semantic information in cognitive architectures. We saw that both the symbolic and the connectionist, horizontal architectures cannot provide basic theory of semantic information for cognition. Semantic information gets presupposed but not explained. Only the embodied approaches have a hope of explaining semantic information as a natural phenomenon. We did not offer such an explanation. One of the most exciting live debates in the philosophy of information has to do exactly with the question of how semantic information emerges in cognition. We saw some of this debase in the chapters on natural information and meaning, and we will re-examine the problem of meaning in the following chapter.

We will end with a short note on a few important omissions that have seen much interest in PI and cognitive science. In this chapter we did not discuss evolutionary accounts of semantic information in cognition. Such accounts attempt to use the idea that biological systems evolve based on mechanisms of natural selection. Such evolution allows formulation of teleological explanations of biological (including cognitive) organization, that is,

[33] Such measures are important for harmony theory. In neuroscience, such measures have been proposed for investigation of cognitive phenomena since the 1960s (Barlow, 1961).

explanations where parts of the systems are regarded as having a *proper function* for the organism. For example, the proper function of the heart is to pump blood (and not to make pumping noises.) We can say this because the heart was selected for this function (and not the other). Some philosophers have suggested that this idea can account for the phenomenon of semantic information, where the proper function is to represent.

We also did not discuss the role of information in the empirical and theoretical studies of modern neuroscience. This is a complex and fascinating subject that demands a chapter of its own. Stay tuned for future editions of this book for such a chapter.

10.7 Exercises

1. Research how bees navigate their environment. What aspect of bees' behaviour can be described in informational terms? Do you think that such a description adds something useful?

2. What do you think are the most important differences between human and animal cognition? Are the most obvious differences also the most important? Which of the identified differences are informational in nature?

3. Imagine a simple example of a chess-playing machine. What would make you say that the information processing by the system contains some meaning? What would make you suspicious that it contains meaning?

10.8 Further reading

For a simple introduction to philosophical problems of cognition see Andy Clark (2001). For important readings on architectural and representational debates about cognition John Haugeland (1997).

Chapter

11

11. MIND

Consciousness, intentionality, and information processing

11.1 Introduction

> *'Consciousness is the biggest mystery. It may be the largest outstanding obstacle in our quest for a scientific understanding of the universe…It still seems utterly mysterious that the causation of behavior should be accompanied by a subjective inner life. We have good reason to believe that consciousness arises from physical systems such as brains, but we have little idea how it arises or why it exists at all. How could a physical system such as a brain also be an experiencer? Why should there be something it is like to be such a system? Present day scientific theories hardly touch the really difficult questions about consciousness. We do not just lack a detailed theory; we are entirely in the dark about how consciousness fits into the natural order.' (Chalmers, 1996)*

Philosophy of mind has historically dealt with, roughly, three issues: rationality, intentionality, and consciousness. The problem of rationality focuses on what it is to be intelligent and to behave rationally. A great deal of progress has been made in our understanding of rationality by studying cognition, the topic of the previous chapter. So in this chapter we will turn our gaze towards intentionality and consciousness (although as we'll see, the three are often interconnected).

The problem of intentionality focuses on how our concepts come to be "about" the things they represent, how words are associated with meanings, and what it is for us to come to have an understanding of those meanings. The problem of consciousness is the most difficult of the three, perhaps because it is so difficult to even say what it is. Consciousness is often described as a kind of awareness or higher-order sense of one's identity and perceptual or mental experiences. Another way to put it is that our experiences have a certain qualitative, subjective, or personal aspect – they have a "phenomenology". Nagel (1974) most famously described our consciousness of being in a state of mind as the experience of "what it is like to be" in that state.

In this chapter, we will cover three representative examples of how the information turn has influenced debates in philosophy of mind. The first example focuses mainly on intentionality and is probably the most famous. It is known as Searle's Chinese Room Argument. Searle (1980) develops a thought experiment in which he purports to show that how we understand the meanings of symbols or words – what we will call semantic understanding – cannot simply amount to the processing of information (where information processing is

understood in terms of a horizontal architecture, i.e., input – processing – output). Although the thought experiment received widespread attention and is still taught in philosophy of mind courses today, it rests on an impoverished understanding of information processing (even given the architecture Searle describes). So one purpose of presenting this example is to correct this by helping the reader understand why this pervasive argument doesn't work, and we will do so by actually appealing to aspects of information processing itself.

In the second example, we will look at how information processing has been used to formulate a view about what consciousness is. One of the earliest standard views is the computational theory of consciousness, which claims that conscious experiences, like the pain of a toothache or the pleasure of a massage, are just instances of the right kind of algorithm being implemented on the right kind of input. We will look at an argument that suggests this information-processing view of consciousness (the computational view) is wrong. The purpose of presenting this example is to expose the reader to one of the sharpest criticisms that consciousness is mere computation.

The third example turns our attention to a more recent development in the study of consciousness. The main issue is about whether – and how – you can know that you are not a zombie. We don't mean whether you are not one of the "undead" as portrayed in popular culture – of course you're not that kind of zombie. What we mean by zombie here is a philosophical zombie: a human that is functionally and behaviourally equivalent to you or I, but has no consciousness. One of the main philosophical issues about zombies is whether you can know you aren't one of them, and in particular, how you would know that. How do you know that, unlike zombies, you are conscious of things? We will look at an argument by Floridi (2005a) which says that, contra Dretske (2003), it is possible to explain how we know this.

One last point before we proceed. Philosophy of mind has been dramatically impacted by the computational notion of information processing. This is because it allowed the introduction of questions that could not have been asked before. We can now ask how concepts that are precisely defined in the theory of computation, like the concept of an algorithm, relate to questions about different aspects of the mind (be it rationality, intentionality, or consciousness). That's not to say, however, that there haven't been nuanced developments in our understanding of information processing, especially when it comes to the relevance of cognitive architecture (see Chapter 10). The point is just that the computational notion of information processing is a tried and true point of departure for tracking various dialectics in the philosophy of mind.

With that said, the reader is asked to have some basic familiarity with the theory of computation, which can be found in Chapter 10. Although it is not required, it is also helpful to have some idea how algorithms would work by manipulating strings.

11.2 Semantic understanding

Consider the following story:

> "Alice went golfing on a hot sunny day in California. She was wearing shorts and a tank top, but to her dismay, she forgot to bring sunblock. At the end of the day, Alice went to the store to pick up some lotion with aloe vera to help relieve some of the pain. In a few days, the back of her neck and her shoulders began to peel."

Did Alice get sunburnt while golfing? The natural answer is "yes". But how do we come to this answer? If you go back to the story, you'll notice that it does not explicitly say that Alice got sunburnt. We get the answer by

"reading between the lines". That is, we use our understanding of the meaning of phrases like "hot sunny day", "tank top", "forgot to bring sunblock", "aloe vera", "began to peel", etc., to infer that Alice must have got sunburnt. This inference is so natural and easy to us that we hardly notice we do it. Our ability to make such inferences is attributed to the fact that we have semantic understanding i.e. we understand the meanings of words and sentences.

Despite such ease, however, we do not have a clear understanding of how we make such inferences. Given the information in the story, the inference appears to be next to trivial to us, but it turns out that building information-processing machines to make the same inference is non-trivial.

One of the earliest develops in artificial intelligence (AI) research was to build programs that can "read between the lines". Roger Schank and his associates for example, built a program called SAM, which stands for Script Applier Mechanism (Schank & Abelson, 1977). SAM is given a story like the one above and then is asked to answer some comprehension questions. The idea behind SAM is that it can extract information that is not explicitly contained in stories by following scripts – data structures that contain stereotypical information about a situation. For example, SAM might have a restaurant script that divides into more specific scripts called tracks: the fast food track, the buffet track, the French restaurant track, etc. The French restaurant track may be divided into scenes such as entering, waiting to be seated, being seated, selecting drinks, etc. SAM can then "read between the lines" when asked questions about going to restaurants because scripts help fill in missing details.

What is interesting about how SAM "reads between the lines" is that it does so just by manipulating strings of symbols. SAM does not understand what the letters, punctuation marks, and words are supposed to represent. SAM is not programmed to know how to associate a string of symbols, like the ones that make up the words "teacup" or "tea", to the things that are referred to by the string of symbols i.e. teacups and tea as we find them in the world (notice that we used quote marks to indicate we mean the word or symbol, not what the word refers to). What SAM is programmed to know is how to associate a string of symbols, again like the string of symbols "teacups", to other strings of symbols like "are objects that can hold tea". In other words, SAM is a syntactic machine. It is programmed to recognize and make inferences about strings of symbols at the syntactic level. SAM is not programmed with semantics, the meanings of certain strings of symbols like "teacup". Nevertheless, despite being just a syntactic machine, SAM appears to exhibit a great deal of behaviour that we tend to associate with semantic understanding.

So a natural question that arises is whether genuine semantic understanding can be exhausted by the right kind of syntactic manipulations like that of SAM. (The reader may recall that a brief discussion about the relation between the syntactic and semantic level was had in section 10.5.) It seems plausible to think that our understanding of a story is explained by the claim that we also have SAM-like scripts and that our minds are algorithms that make use of them on a syntactic level. It's this kind of thinking that inspired John Searle to develop an argument that no algorithm, no matter how sophisticated its capacities in manipulating syntax, will amount to semantic understanding. We turn to that argument now.

11.3 Searle's Chinese room

To appreciate the argument that semantic understanding is not just the running of the right program, we shall briefly discuss the kind of architecture on which programs like SAM can be based. This architecture we describe is associated with the fathers of modern day computing: Alan Turing and John von Neumann.

We can think of SAM, and just about any digital computer, as having three parts: executive unit, store, and control. The executive unit is nowadays often called the central processing unit (CPU) and it is the part that does the actual work: it reads and writes from storage space. In SAM, the executive unit is the microprocessor in the computer where the SAM program is running. In the storage space we find the other two parts, the store and the control. The store contains the data structures that the executive unit manipulates and updates. Today, we think of files as being part of the store. In SAM, the store contains the scripts, the story under consideration, and the latest input question. The control is the program that guides the behaviour of the machine. In modern terms, the control consists of the applications that are currently running. In SAM, the control is the list of instructions that direct the machine to interact with the scripts and input in a certain way that eventually leads to the output answer.

Using this kind of architecture, the philosopher John Searle has developed an argument for why programs like SAM cannot be true understanders (Searle, 1980). He has us imagine the following kind of system, which he calls "The Chinese Room". Here, we have a large closed room with two slots, one for input and the other for output (notice that this is the horizontal architecture we discussed at the beginning of this chapter and in Chapter 10). The system has three components like the ones we described above. John Searle is in the room and acts as the executive unit. It will be an important fact that John Searle does not speak Chinese. The store consists of a warehouse of scripts written in Chinese, a story written in Chinese, and an input question that is also written in Chinese. Lastly, the system also has the control, which is a list of instructions written in English that tell Searle how to treat the structures in the store.

The system begins to run when a story is input into the store. Searle, who lives inside the system, effectively implements the SAM program by a process of pattern matching. He does not understand any of the Chinese characters in the input, but the instructions, which are written in English, tell him what to do with them. By following the instructions, Searle eventually writes down a string of Chinese characters that the instructions tell him to send through the output. Searle, not knowing any Chinese, has no idea that he is expanding a story that happens to be written in Chinese. To him, he is just boringly manipulating characters that he does not understand.

Searle (the real one, not the one we are imagining in the machine!) claims that no understanding occurs when the Chinese Room implements SAM. The executive control (the imagined Searle in the room) does not understand what he is doing. And surely, it seems ridiculous to say that the instructions and the books of Chinese characters add understanding – they are just ink marks! In fact, it doesn't matter whether the Chinese Room is implementing SAM or some other program. Searle's point is that because any program can be implemented in this manner (this happens to be a provable fact about the kind of computer architecture we've described, which we discuss below), and because no understanding whatsoever occurs in the room, no genuine understanding can be had in virtue of running a program.

To be clear, the point of the argument is not to show that the kind of information processing in the Chinese Room is not important to semantic understanding. In fact, Searle thinks that humans are a kind of computer, and so since humans are intelligent and have semantic understanding, some computers do too. Searle's point is

that intelligence is not the result of merely instantiating the right computer program – he contends there must be more to it than that. The Chinese Room, and computer programs in general, are simply doing syntactic symbol manipulation: there is no mastery of semantics or understanding of what the symbols mean.

11.4 Response to Searle

There have been numerous responses to Searle's line of argument (you will find further reading at the end of this chapter). We shall discuss a particular version of a response which says that, although no component of the Chinese Room has semantic understanding, the room as a whole does (also known as the systems response). In particular, we will briefly discuss what has been called the virtual machine response.

The virtual machine response is motivated by the very sort of architecture we've covered above. Recall that the storage space is divided into two parts, the control and the store. The control can consist of a particular program that runs on the input data in the store. We could in principle build a machine that does not have an executive unit (CPU), but such a machine would only be able to run that one program on the given input. By introducing an executive unit, we can build a computer that can simulate any other computer whatsoever! For this reason, we call them universal machines. Your modern day computer is a universal machine: you just need to give it a description of the program you want it to run, and it can simulate it for you – no need for you to buy a new machine every time you want a new program! The program that is being run by the universal machine is referred to as the virtual machine for exactly this reason.

This feature of computers leads to an interesting and relevant fact: a universal machine will be in two states at any one time. One of the states will be of the universal machine as a whole; the second state is that of the machine it is simulating i.e. the state of the program that it is running. For example, the universal machine might be in state that tells the executive unit what to do next, given the state of the virtual machine. Why is this interesting? It turns out that some things that are true of the universal machine are not true of the virtual machine that is being run. For example, what counts as input is different for the two machines. For the universal machine, the entire storage space (which includes both the store and the control) counts as input. For the virtual machine, however, only the store counts as input – the data in the control is, in a sense, the virtual machine. So both the universal and the virtual machine are on the same physical device, but they have different inputs. For example, the virtual machine might be an adding machine. It takes as input two numbers in the store, and then returns the sum of those two numbers. The corresponding universal machine however, is not an adding machine. Its input is not just two numbers, but also the code for the adding program, which it then simulates. So there are different truths about these two machines.

Why is this fact about universal machines interesting? Since we imagined the Chinese Room to have the same kind of architecture as the computing machines we discussed (it has an executive unit, a store, and a control), the same lessons about the universal machine transfer over to the Chinese room. In fact, in the exact same way that a universal machine runs a virtual machine, the Chinese Room runs a virtual machine: a Chinese speaker! Although the input to Searle (the one in the room) is a list of instructions that tell him what to do with Chinese symbols that he does not understand, the input to the Chinese Room is entirely different. The inputs to the Chinese Room are Chinese questions.

To make the point even clearer, imagine that the virtual machine is called Soo Lin. If you ask the Chinese Room what its name is, it will respond with, "My name is Soo Lin." Searle, who is playing the role of the executive unit, would never say his name is Soo Lin. For one, his name is not Soo Lin! But more importantly, this is because Searle is never asked what his name is. This is because, as we just pointed out, the input to Searle

is different to the input to the Chinese Room. Moreover, it is not just the inputs that are different, but the outputs as well. Searle put together a string of Chinese characters that he does not understand, but Soo Lin, the Chinese speaker that is the virtual machine, does understand Chinese.

We can turn this around. Suppose we were to ask Soo Lin a question in English, a language which Soo Lin does not understand. Searle obviously does understand English, and so he will understand the input. However, because Searle is the executive unit, he will look up in the instructions what he should do with English questions. The instructions will tell him to write down some Chinese characters and then output them. What Searle doesn't know, since he doesn't understand Chinese, is that he has written down a message that says, in Chinese, "I'm sorry, I do not understand English". That's exactly the kind of response we would expect from Soo Lin!

In sum, the point of the Virtual Machine Response is that the Chinese Room argument rests on an impoverished understanding of computation. It is perfectly fine to admit that no component of the machine has understanding, in the same way that no component of the Chinese Room has understanding. The question about whether there is semantic understanding going on is about the virtual machine, the running of the program. It seems quite plausible that Soo Lin has semantic understanding of Chinese, since "she" will do just as well as any human Chinese speaker would. So it seems that semantic understanding might be the result of running the right program after all.[34]

11.5 Is consciousness the result of information processing?

We have covered an argument by Searle which attempted to show that semantic understanding cannot be just a matter of the right kind of information processing, and we also showed that Searle's argument can't be right if we have a more complete understanding of computation. We will now look at an idea that conscious experience is the result of information processing.

What do we mean by consciousness? Philosophers have in mind what is called phenomenal or experiential consciousness. A person is said to be experientially conscious if there is something that it is like to be them. Our experiences of colour, sounds, the awareness of the thoughts in your head, etc., are examples of experiential consciousness. There is something it is like to have these experiences. In fact, this kind of consciousness is a state you can be in even when you're asleep, for there is something to be like for you to have a dream! However, paper, pencils, and pens, do not have experiential consciousness because there is nothing that it is like to be them.

But what about machines? Could they ever be conscious? This is a difficult question. Maybe if machines run programs that are complex enough, they will be conscious. Then again, it seems that no matter how complex we make a machine, it won't come to have experiences of pain or pleasure, for example. Making things more complex doesn't seem to answer the question.

The philosopher Tim Maudlin has argued that consciousness cannot arise merely in virtue of running the right program, regardless of complexity (Maudlin, 1989). For experiential consciousness to occur, more than mere computation is required. Maudlin's argument has the structure of a *reductio ad absurdum*. That is, Maudlin assumes (for the sake of argument) that consciousness arises by simply running the right program. He then

34 If you are interested in learning more about this debate, have a look at (Copeland, 1993).

goes to show that this assumption leads to absurdity. Hence the assumption must be false. Consciousness is not just the mere running of the right program (no matter how complex).

The argument starts with the claim that experiential consciousness of a human at a particular time is determined entirely by our brain state at that time. It also seems possible for creatures with non-human brains to have experiential consciousness as well. In that case, an instance of their experiential consciousness will be determined entirely by their physical state at that time. But what if two creatures (either human or non-human) have the exact same physical states? Maudlin suggests that when two creatures are going through the same physical states, then they have the same conscious experiences through that period (if they are having conscious experiences at all). Maudlin calls this the Supervenience Principle.

Maudlin suggests that the Supervenience Principle is very plausible. However, when we combine this principle with another condition, we end up with some very strange results. What is this other condition? It is motivated by the *reductio* assumption. Consider an instance of human experiential consciousness, such as your experience of a toothache. The *reductio* assumption says that experiential consciousness is merely a matter of running the right program. That means that your conscious experience of a toothache is a matter of your brain running a "toothache algorithm". More generally, whenever the "toothache algorithm" is being run, whether by you or something else, there is a conscious experience of having a toothache going on. Maudlin calls this the Sufficiency Condition.

To see how the Supervenience Principle and the Sufficiency Condition lead to conflicting results, we need to explain a few more facts about computation. One interesting fact is that computation is medium independent. That means that the running of a program or algorithm is not determined by the medium that is doing the running. The ADD algorithm, for example, can be run by silicon chips arranged in the right way to manipulate electrical inputs (where the values of bits are high and low voltages). But the ADD algorithm isn't limited to silicon chips! We can build machines made out of other things that still implement the ADD algorithm. We could use a mechanical system (made out of metal gears, for example) that uses punchcards to encode the instructions for adding two inputs. Or even better, consider yourself: you are not made out of silicon or metal gears, and yet you can implement the ADD algorithm (and you might have done so by trying the exercise above). The point is that it doesn't matter what a system is made of when we say that it is running an algorithm.

So medium independence in the context of information processing means that it doesn't matter what a system is made of when it comes to running algorithms. But now consider the claim that your conscious experience of a toothache is a matter of running a "toothache algorithm". That means that there is another system, made of something other than the stuff of your brain and body, that can run that same "toothache algorithm". In fact, this other system could be a very complicated arrangement of water troughs! It could be built so that a trough full of water encodes for 1, an empty trough encodes for 0 and the running of an algorithm is a matter of trolley that runs back and forth along a line of troughs, following instructions on how to fill and empty troughs (instead of writing down a 1 or 0, it fills a trough with water by using a hose, or it momentarily tips the trough over to empty it). Let's give the name Klara to this other system that runs the same "toothache algorithm".

The Sufficiency Condition (that a run of the "toothache algorithm" is sufficient for producing a conscious experience of a toothache) now gives us a slightly strange result. Every time Klara runs the "toothache algorithm", Klara will experience a toothache; or better, the virtual entity that Klara generates will experience a toothache (recall our discussion of Soo Lin and Searle above). This is somewhat odd because neither Klara nor the virtual entity have any teeth. Though strange, that doesn't actually matter. We can imagine that a toothless

person could still have conscious experiences of toothache – it is well documented, for example, that people can experience pain in limbs they have lost (known as phantom pain).

Still, it may seem a little strange that a system like Klara can have conscious experiences. After all, it is a system just made of a line of troughs and a trolley that runs back and forth filling and emptying troughs. But strangeness alone is not enough reason to rule out this possibility. Klara's dynamics, the sequence of moves that the trolley makes filling and emptying troughs, might be very intelligent, even though we do not recognize it as such. So to refute our assumption, we need a result that is truly absurd.

To do that, Maudlin builds a very dumb version of Klara through a series of reconstructions. We can skip over the details here.[35] The main idea is that every newly constructed machine will either: i) undergo the same physical activity as the previous machine, and so by the Supervenience Principle will have the same conscious experience as the previous machine; or ii) run the same program as the previous machine, and so by the Sufficiency Condition will have the same conscious experience as the previous machine. The resulting system, called Armature, is simply a series of troughs lined up on a hill from top to bottom and a trolley that rolls down alongside the troughs. Armature is a system that runs the same "toothache algorithm" as Klara, except that the troughs have been rearranged so that for a particular run of the algorithm, no troughs need to be filled or emptied. It is like an instance of running the ADD algorithm where all the appropriate 1s and 0s for the answer have already been written down, but now the 1s and 0s are troughs full or empty of water, and the algorithm that is being run is the "toothache algorithm".

The truly absurd result we get with Armature is that it (or the virtual entity being generated) will experience a toothache just as Klara would, except that, where Klara seems to be some kind of intelligent system, Armature is just a trolley rolling down a hill alongside a series of troughs! Something has gone wrong, but what? At each step of the reconstruction from Klara to Armature, either the Supervenience Principle or the Sufficiency Condition was used. In order to say that Armature is not experiencing a toothache, we must give up one of these. Maudlin suggests that we have to give up the Sufficiency Condition because the Supervenience Principle is more plausible. Giving up the Sufficiency Condition, however, means that conscious experiences like toothaches are not just the result of running the right program. Consciousness, according to Maudlin, is not just the result of information processing.

11.6 How do you know you're conscious?

By the time you finish reading this sentence, you'll be aware that you're conscious. So we'll take it as a given that you know you're conscious. That much seems obvious. But now here is a rather difficult question: *how* do you know that you're conscious? What is it about consciousness that tells you that you are conscious? Is there anything you can be aware of, either externally or internally, that helps you tell that you are not a zombie? After all, a zombie is functionally equivalent to you. It walks and talks just like you. It responds to pains and pleasures just like you. It may even *think* it's conscious just like you. But it isn't.

Fred Dretske (2003) argued that there is no way for you to know that you know you are conscious. Again, the claim isn't that you don't know you are conscious. Of course you know that. What Drestke is trying to figure out is whether there is a way of knowing that, unlike zombies, we are conscious of things. His answer is that there isn't: there is nothing you are aware of that tells you that you are aware of it. So it may be impossible to answer the question "how do you know you are not a zombie?"

35 If you wish to look up the details, they are in the original paper (Maudlin, 1989).

Luciano Floridi (2005a), however, argues that it is possible to explain how it is that you know that you are a conscious agent and not a zombie. To do this he develops a reliable and informative test called *the knowledge game*. There are four different versions of the game, each building on the previous, but we can skip ahead to the fourth version to get the general idea.

The game involves three agents that are offered five pills. Three of these pills don't do anything, but two of them make the agents completely dumb. Now suppose the first agent, Bob, takes one of the pills. A few moments later Bob is then asked which pill he took, i.e., whether he is in a dumb state or a non-dumb state. Is there a way for Bob to know this?

We can rule out a few things. First, in line with Drestke, we're ruling out that Bob has some special kind of access to his own mental states – what Bob is aware of does not come with a special internal label that tells him what he's aware of. So Bob can't come to know which pill he took that way. Second, Bob can't know which pill he took based on externally inferable information – all the pills look the same. And no one else has taken a pill yet, so he can't rule out the possibility that the other two agents took the dumb pills (at least then he could see that they suddenly went dumb and infer he took an innocuous pill). Nor can Bob know which pill he took by some kind of "bootstrapping" i.e. the pills don't have any properties other than making an agent dumb, so they don't make Bob feel any different than an innocuous pill would.

In short, there is no way for Bob to know or infer that he is in a dumb state. But now Bob can make at least one inference. Since he cannot rule out the possibility that he is in a dumb state, Bob will say that he is ignorant about which state he is in. For all he knows at that moment, he might be in a dumb state or he might not be. Now after Bob makes this inference, there are two things that could happen. Either Bob's report fails to trigger any further reaction in himself, or his report spawns a series of counterfactual reasoning. If Bob's report doesn't trigger any further reactions, then Bob has failed the test and thereby lost the knowledge game. Bob remains in his state of ignorance about whether he is in a dumb state or not.

In the second case, Bob hears his own report that he is in a state of ignorance. This leads him to the following kind of reasoning: "If I'd taken the pill that makes me dumb, I wouldn't have been able to report that I was ignorant about which state I'm in. But I did report it and I know that I did because I heard myself speak and saw the guard acknowledge my report. The only way I could have reported my ignorance is if I hadn't taken the dumbing pill. So, now I know that I can't be in a dumb state, which means I have to be in a non-dumb state. The only way I'm in a non-dumb state is if I took a non-dumbing pill, so I know that I took the non-dumbing pill. Since I know all of this, I can go back and revise my previous report that I am in a state of ignorance, because I am no longer in a state of ignorance and I know this. Moreover, by having gone through this whole process of reasoning and passing the test, I can also report, correctly, that I am in a state of knowing *how* I know *both* that I didn't take the dumbing pill, *and* that I *know* I didn't take the dumbing pill." In sum, shortly after his report on his state of ignorance, Bob corrects himself and reports, accurately, that he is not in a dumb state (and that he knows this, and that he knows how he knows this).

Let's think about what it takes for Bob to be able to correct himself and thereby pass the test. Bob has to be able to identify that it was *he* who was the agent reporting *his* state of ignorance of *his* state in question, and that it was *he* who was playing the game and that the guard asked *him* the question after *he* took the pill. So Bob can pass the test only if he recognizes that it is himself in the game. The same is true for any agent that plays the game: they have to be able to capable of this *subjective reflection*.

Why does passing the test turn on the capacity of subjective reflection? Because that is a hallmark difference between zombies and humans. Subjective reflection not only requires the capability of understanding semantics – something that zombies also have – but it also requires consciousness. An agent that isn't conscious doesn't have an experience of what it is like to be in a certain mental state, and isn't aware of its own personal identity and mental experiences. If Bob were a zombie, he wouldn't be conscious, which means he wouldn't have the capacities needed to correct himself after reporting his state of ignorance, which means he would fail the test. So passing the test – and thereby winning the game – means you have to be conscious.[36]

In sum, by developing the knowledge game, Floridi has argued that, contra Dretske, there is a possible way of knowing that we aren't zombies. Of course, there may not be such dumbing pills around for us to use (and they probably wouldn't pass any ethics boards anyway). But that's not the point. Rather, what we wanted to figure out is whether there is *any* way of telling the difference between zombies and conscious agents like ourselves. The knowledge game demonstrates that there is a possible way of obtaining such information.

11.7 Conclusion

We considered three examples where the notion of information has impacted developments in philosophy of mind, particularly with respect to intentionality and consciousness. Searle's Chinese Room thought experiment attempts to establish the claim that semantic understanding cannot be merely a matter of running the right program. But we showed that Searle's argument is based on an impoverished understanding of computational information processing. This example clearly demonstrates how philosophy of information enriches our understanding of the mind. Here's another way to think about it: from the perspective of information, the criterion for something to exist is not that it has to be immutable, as a physicalist metaphysics might suggest. Rather, existent things are potentially subject to interaction, even if they are intangible! (Floridi, 2011c). That is precisely what the response to Searle highlights. When one presents questions to the room in Chinese and receives answers in return, one is not interacting with Searle (who is inside the room), or the books in the room, or the system as a whole. Rather, one is interacting with a virtual entity, Soo Lin, and it is the virtual entity that is answering the questions. The fact that this is a viable solution to a difficult problem in the philosophy of mind is testament to the fruitfulness of the perspective of philosophy of information.

In the second example we looked at Maudlin's argument that conscious experiences are not just the implementation of a "toothache" algorithm (or some other "experience" algorithm). The way to understand what is meant by "the running of a program" or "the implementation of an algorithm" in this example is in terms of a kind of information processing that is computational. So one might ask whether something like our response to Searle would work in the case of Maudlin. The fact that Klara behaves like an intelligent system might suggest that there is a virtual mind having an experience of a toothache. But then again, we can build Armature, a system just like Klara, except all it does is roll down a hill alongside a series of troughs. It is hard to see how a virtual mind could be generated by simulation if there is nothing intelligent happening in the system. So the problem of understanding conscious experiences continues to be an open one.

That said, the third example, Floridi's knowledge game, demonstrates that it is possible to explain how we know that we are conscious and not zombies. This counters the pessimism of Dretske, who thought that there is no way for us to know how we know that we are conscious. Floridi's knowledge game gives us such an informative test, even if such a test is difficult to carry out in the real world.

36 Note that you might fail the test for other reasons besides not being conscious, e.g. you might not have the means to speak. So failing the test doesn't mean you aren't conscious. The knowledge game is a sufficiency test (much like Turing's Imitation Game): winning the game means you have the feature in question – consciousness – but losing the game doesn't tell you anything.

11.8 Exercises

1. Searle's argument uses a particular kind of computational architecture consisting of an executive unit (CPU), a store, and a control. In the chapter on cognition, we considered other architectures of information processing. How much does Searle's argument rest on the particular architecture that he picked? Could he try to make the same argument using a different architecture? And would that avoid the objection we gave, or would it still be pretty much the same?
2. Suggest how one might respond to Maudlin's argument, particularly the "truly absurd result" we presented at the end. Hint: in the final remarks we suggested that it is "hard to see" how a virtual mind could be generated when nothing intelligent appears to be happening in a system (i.e. no physical changes are taking place). Might there be a reason to doubt our intuitions here, given that the "intelligence" of Klara resides in something other than physical activity?
3. Floridi's knowledge game turns on the notion of subjective reflection. This means he has a certain conception of what it means to be conscious: "a state in which the agent and the I merge and 'see each other' as the same subject." Can you think of other convincing ways of characterizing consciousness that do not require the notion of subjective reflection, or does it seem like this is an essential feature of any characterization of consciousness? For example, is it possible for an agent to experience the painfulness of pain without recognizing that it is *her* pain?

11.9 Further reading

Copeland (1993) is a great introductory text that goes into more detail than this chapter. The original paper (Maudlin, 1989) is significantly more difficult, but will reward study. Penrose (1989) is a very interesting text that still inspires controversy. For a comprehensive overview of the computational approach to rationality, intentionality, and consciousness, see Rey (1997). Chalmers (1996) is a good way into the debates surrounding consciousness and zombies.

Part V: Formal foundations

Chapter

12

12. Logic

Logic in PI and information for logic

12.1 Introduction

> *'Logicians have apparently failed to relate their subject to the most pervasive and potentially most important concept of information.' (Hintikka, 1973).*

When we study logic, our primary aim is to separate good arguments from bad arguments. If, for instance, Alice promises that if Bob does the job he will get a reward, Bob can, once he has completed the job, argue that he should get his reward. In that case, Bob used a good argument form called *modus ponens*. Similarly, if Carol knew about Alice's promise, but also heard Bob complain that he didn't get a reward, she could conclude that Bob didn't do the job. Here, Carol used the good argument form called *modus tollens*. Yet, if Carol had heard Bob brag about not having done his job, and thus concluded that Bob didn't get a reward, her reasoning would have been fallacious. She would have used the bad argument or fallacy called *denying the antecedent*.

But why is this last argument not acceptable, and what makes good arguments good? Here, several answers are possible. One answer is that good arguments are just those arguments whose conclusions cannot be false whenever their premises are true, whereas bad arguments leave open the possibility of the conclusion being false. Thus, for instance, Bob's failure to complete his job together with Alice's promise that he would get a reward for doing the job doesn't make it impossible for Bob to get a reward for some entirely different reason.

The view that arguments are good because (and only because) the truth of their conclusion is ensured by the truth of their premises is the primary (but not the only) criterion we use to evaluate arguments. Equally often we will say that an argument is good because the information in the conclusion of that argument was already part of the information in the premises. A good argument is an argument whose conclusion does not "go beyond" its premises. Both criteria hint at an important feature of good arguments. The truth-based criterion emphasises the fact that if we only reason with good arguments, we will never step from true premises to false conclusions; the information-based criterion emphasises the fact that we can use arguments to extract information from what we already accept.

Within the philosophy of information, logic can both be a topic and a part of our methodological toolbox. In this chapter we will highlight both these aspects. After a brief prelude in which the key technical notions are explained, we first relate logic to the three core methods of the philosophy of information (minimalism,

constructionism and levels of abstraction), and then take a closer look at the previously introduced idea of logic as a tool for information-extraction.

12.2 Prelude

This prelude aims to introduce the minimum amount of formal logic that is required for getting through this chapter. It is built around the introduction of a number of key concepts in formal logic.[37]

By a formal language, we mean a schematic language that is introduced by, first, specifying what the logical and non-logical symbols of our language are, and, second, by enumerating the different ways in which these symbols can be put together (so-called formation-rules) to obtain the set of all the admissible expressions of our language. We often refer to these expressions as well-formed formulae.

Example 1: The language of propositional logic. In a propositional language, the only non-logical symbols are atomic propositions of the form *p, q, r* …, while the standard logical symbols of propositional logic include:[38] *and*, *or*, *implies*, and *not*. Using these building blocks, we say that: (i) all atomic propositions are well-formed formulae; (ii) if *A* and *B* are well-formed formulae, then *A and B*, *A or B*, and *A implies B* (sometimes written as *if A then B*) are well-formed formulae; (iii) if *A* is a well-formed formula, then *not-A* is also a well-formed formula; and (iv) nothing else is a well-formed formula.

Once we have these guidelines, we can always find out whether or not a given string of logical and non-logical symbols is a well-formed formula of the language of propositional logic. This is the language that allows us to say things like "*A and B*" or "*(A and B) implies C*". Specifically, we could use a propositional language to express our example from the introduction, for the promise made by Alice has the form "Bob does the job *implies* Bob gets a reward" where both "Bob does the job" and "Bob gets a reward" are atomic propositions.

Example 2: The language of first-order logic. A predicate language relies on more than one type of non-logical symbols, namely predicates and relations of the form *P, Q, R* …, variables of the form *x, y, z* …, and constants of the form *a, b, c* …. For the logical symbols, we only need to add the quantifiers *All* and *Some* to the logical symbols we already introduced for the propositional language. Using these building blocks, we say that: (i) when *P* is a predicate and α is either a constant or a variable, then $P\alpha$ is an atomic formula;[39] (ii) when *P* is a predicate, *x* a variable, and *A(x)* a formula that may or may not contain *x*, then *Forall x A(x)* and *Some x A(x)* are also well-formed formulae; and (iii) if *A* and *B* are well-formed formulae, then *A and B*, *A or B*, and *A implies B* are well-formed formulae; (iv) if *A* is a well-formed formula, then *not-A* is also a well-formed formula; and (v) nothing else is a well-formed formula.

The language of first-order logic allows us to say things we couldn't yet say in our purely propositional language. We can, for instance, say that "All computers have processors," (*Forall x (Cx implies Px)*) or that "Some computers are fast" *(Some x (Cx and Fx))*.

Once we have a formal language, we can exploit the precise way in which we defined that language to do at least three things. First, we can say what it means for a formula of a given language to be true (or false); second, we can say what it means to be able to obtain (or deduce) one formula from one or more other formulae; third, we can use features about truth or about deducibility to give a general account of what it means for some

[37] Readers who would like to know more are encouraged to take a look at the further reading section at the end of this chapter.
[38] It is customary to use symbols to refer to the logical symbols. Here, we stick to the natural language counterpart of these symbols.
[39] Generally, when *R* is an *n*-ary relation, $Ra_1 \ldots a_n$ is also an atomic formula.

formula A to follow from a (possibly empty) set of formulae B_1 …, B_n. Formulae that are deducible from an empty set of formulae are called *theorems*; formulae that are always true are said to be *valid*.

When we say that logic explains what it means for a formula to be true or false, we do not mean that logic alone can help you to determine whether or not a certain formula is true. Instead, it will show you how, given the structure of a language, the truth or falsity of a formula can be reduced to the truth or falsity of the components of that formula. To that end, logicians give a definition of *truth-in-a-case* that exploits the systematic structure of a formal language by, for instance, stipulating that "*A and B"* is true in a case *c* if and only if "*A"* is true in *c* and "*B"* is true in *c,* or by stipulating that "*not-A"* is true in *c* if and only if "*A"* is false in *c*. Such rules can then be used to systematically reduce the question of the truth of a formula A to questions concerning the truth or falsity of the atomic formulae that occur in A. Finally, the truth of an atomic formula in a case is something that cannot be further reduced. This means that, at least for our purposes, we can think of *cases* as things that decide which atomic formulae are true, and which are false.

In many ways, the existence and construction of proofs form the focal point of logic. By a proof, we mean a set T of formulae B_1 … B_n that (a) are organised in a list, tree, or other type of structure, where (b) a possibly empty subset S of T is taken to be given (the *premises* of the proof); (c) all other formulae in T are obtained by applying certain rules to the formulae that are also in that list; and (d) a single formula A in T is called the conclusion. One standard form for a proof is just an ordered list of formulae, where the first *n* formulae are the premises, all other formulae are obtained by applying rules to the formulae higher up in the list, and the final formula is the conclusion. Here's an example based on our Bob and Alice example:

(1) Bob does the job implies Bob gets a reward (premise)

(2) Bob does the job (premise)

(3) Bob gets a reward (concluded from lines 1 and 2 by *modus ponens*)

Whenever we have such a list, we say that A can be deduced from S. Furthermore, when this S is empty, we will say that A is a theorem (it can be deduced from zero premises).

Obviously, if we use sensible rules to deduce A from S, having a proof will be sufficient to judge that A follows from S. Another way to think about what it means for A to follow from S refers to the notion of truth: if all premises in S are true, then A will be true as well. This analysis of *follows from* is usually called *truth-preservation*, for it guarantees that the truth of the premises carries over to the truth of the conclusion. Given our prior analysis of truth, we can make this idea of truth-preservation more precise by stipulating that A follows from S if and only if:

(Cases) Every case where all members of S are true is also a case where A is true.

Here too, when S is empty, we say that A is valid (it is true in all cases). Both our accounts of "follows from" are meant to capture the same connection between premises and conclusions. Specifically, we care about the following two features. First, we want our proofs to be *sound,* which means they have to preserve truth; second, we want our rules for the construction of a proof to be *complete*, which means they have to capture all instances where truth is preserved.

Logicians do care a lot about soundness and completeness, and spend a lot of time proving that certain systems indeed have these properties. Fortunately, the details of soundness and completeness proofs belong in an advanced logic course, and are well beyond the scope of this chapter. We will mainly be concerned with the fact that by being more or less restrictive about what counts as a *case*, we can tell several different stories about what it means to follow from. Basically, the following inverse relationship holds: if we are fairly restrictive about what counts as a case, more things will follow from our premises, whereas if we are quite liberal about what counts as a case, it will be easier to find a case where all the premises are true and yet a certain conclusion is false.

The traditional way of delimiting what counts as a case is based on the following criterion: c is an admissible case if and only if for every atomic formula A (of a propositional or first-order language) c makes A either true or false, but not both. When all and only such cases are taken into account, the resulting logic is called classical (propositional or first-order) logic. Quite often, we will say that classical logic only takes into account cases that consistently decide every issue. In almost all cases, non-classical logics are obtained by relaxing this demand. That is, by allowing for cases that leave the truth or falsity of some (atomic or other) formulae undecided, by allowing for cases that make some (atomic or other) formulae both true *and* false, or even by dropping both the requirements of exhaustiveness (everything is either true or false) and exclusiveness (nothing is both true and false).

12.3 Logic in the Philosophy of Information

As explained in the introductory chapters, the philosophy of information (PI) isn't just defined by its subject-matter, the nature and dynamics of information, but also by its method, namely the use of information-theoretic and computational methodologies. A substantial part of this methodology, including the method of levels of abstraction that was described in Chapter 2, is inherited from the use of models in scientific practice, and from an area of theoretical computer science called Formal Methods. Logical methods are often used by philosophers of science and theoretical computer scientists, and so we can expect that the same logical tools will also be part of the toolbox of the philosopher of information.

The main purpose of this section is to illustrate how logical methods can be put to work within the philosophy of information. It is structured around the three core methods of the philosophy of information: minimalism, constructionism, and the method of levels of abstraction. For each methodology we shall highlight how it is related to logic and the use of logical tools.

12.3a Minimalism

The method of minimalism (Floridi, 2011b; Greco, Paronitti, Turilli, & Floridi, 2005) deals with the common difficulty that philosophical problems are rarely independent of each other, and that the answer to almost any philosophical question presupposes answers to other outstanding questions. For instance, problems in epistemology may presuppose answers to several questions in metaphysics or in the philosophy of language, which in their turn may presuppose answers to questions in logic or formal semantics, which again may or may not depend on certain metaphysical issues, etc. When the solution to a certain problem overtly or tacitly relies on the existence of answers to other problems, the proposed solution is radically weakened. It is only good if the answer it relies on also works. Minimalism aims to provide stronger answers by relying on fewer outstanding questions. This leads to the following methodological maxim:

(Minimalism) Do not presuppose the answers you do not have.

As a method, the role of minimalism is to provide criteria that allow us to choose tractable problems and to keep track of the interdependence between several sub-problems. As such, the goal of minimalism is closely related to what we try to achieve with the method of levels of abstraction (questions are always formulated at a certain level of abstraction), and with constructionism (good answers are conceptual artefacts or models).

Minimalism also explains the importance of logical methods in the philosophy of information, and this is why we begin with it here. On a traditional conception of logic, we can use logical methods and tools without thereby presupposing the existence of answers to other problems. Logic, in other words, is a safe starting point, and relying on it is consistent with minimalism. This is especially true when it comes to fundamental principles of logic, like the law of non-contradiction (no contradiction is true). Such laws, according to David Lewis (Lewis, 2004), are among the most certain principles that can be assumed in a debate: they constitute a common ground that one shouldn't give up lightly.

Within the philosophy of information, the law of non-contradiction plays a similar role. This will become clear when we relate consistency to constructability, and when we take a closer look at the role of consistency in the method of levels of abstractions.

The use of logic isn't merely consistent with minimalism: it is actually presupposed by the method of minimalism. Minimalism is about keeping track of commitments, and understanding the relation between different problems within a broader problem-space. Such relations are inferential, and are therefore logical and/or probabilistic relations. As a consequence, we cannot select tractable problems without already presupposing a minimum of logic and probability theory.

In summary, because logic (but also probability theory and perhaps some more mathematics) is both consistent with minimalism, and presupposed by minimalism, it plays a crucial role within the philosophy of information and its methodology.[40]

12.3b Constructionism

Constructionism is both a general philosophical attitude that emphasises the maker's knowledge tradition (see Chapter 9), and a philosophical methodology. The philosophy of information, as put forward in Floridi (2010c), emphasises constructionism in both these senses: it sees philosophy as conceptual engineering and argues that theories need to be tested by implementing them in a conceptual model. Here too, we can formulate a methodological maxim:

(Construction) Only rely on what you can actually build.

It is at the stage of the implementation that logic comes into play. We highlight the role of logic in conceptual engineering by contrasting it with its role in conceptual analysis.

The purpose of conceptual analysis is, according to a highly influential conception of what this means,[41] to make sense of our everyday thought or to provide cleaned-up versions of our best folk theories. That is, the process of conceptual analysis should not create novel concepts, but should merely elucidate existing concepts (Beaney, 2012). Conceptual engineering, by contrast, is 'engaged with creating, refining, and fitting together our

[40] While this may suggest that PI is biased towards classical propositional and first-order logic, this need not be so. It is only claimed that minimalism presupposes the use of logic, but this may very well be some non-classical logic.
[41] The so-called Canberra Plan put forward by Frank Jackson and David Lewis.

conceptual artefacts in order to answer open questions' (Floridi, 2011b, p. 293), and is therefore not constrained by pre-existing concepts.

Logic is essential to both conceptual analysis (especially when understood as logical analysis) and conceptual engineering. This is, first, because we want to eliminate imprecisions (and fill in the gaps) of our informal concepts; and, second, because we want to arrive at something coherent. With regard to the aim of precision, the difference between conceptual analysis and conceptual engineering is negligible. With regard to the aim of coherence, the difference between them is more marked. In the former case, the result of analysis needs to be logically coherent or consistent because if we assume that the existing concepts are essentially correct or rational then—as per the principle of charity—its analysis should at least be consistent, and thus comply with the laws of logic.[42] In the latter case, since the construction of concepts does not presuppose pre-existing concepts, there is no comparable assumption that the concepts we start from are already on the right track. Here, the need to come up with logically coherent concepts has a different source, namely the need to obtain conceptual artefacts that function correctly. In other words, because inconsistency is just a form of malfunctioning, logically incoherent artefacts are unacceptable; they do not deliver what they promise (Floridi, 2010d).

Does this also mean that constructionism, or conceptual analysis for that matter, is committed to the canons of classical logic? Here, the answers may diverge, and one way to make this clear is in terms of two notions of consistency, namely absolute consistency and negation consistency.

A set of formulae *S* is negation consistent if and only if there is no formula *A* such that both *A* and not-*A* follow from *S*; it is absolutely consistent if and only if there is at least one formula *A* that doesn't follow from *S*. Conversely (and more intuitively), *S* is absolutely *in*consistent if and only if everything follows from it (we call such sets trivial, for they make everything trivially true, including such absurdities as "the moon is made of green cheese"), and simply *in*consistent if and only if it contradicts itself.

Because in classical logic any formula follows from a contradiction,[43] both notions of consistency (and inconsistency) are classically equivalent. This isn't so for paraconsistent logics, where negation-inconsistency does not collapse into triviality. The disagreement between classical and paraconsistent logic can be explained as follows: according to classical logic, every contradiction (i.e. every expression of the form *p and not-p*) is absurd and therefore necessarily false. While paraconsistent logicians agree that all absurditities are necessarily false, they will also claim that some contradictions are not absurd. A particular type of paraconsistent logician, namely the *dialetheist,* will even make the stronger claim that some contradictions are true (and false as well).

Whether it is the result of conceptual analysis or of conceptual engineering, a theory that is trivial is clearly unacceptable. It cannot be rational (Priest, 2006), and can't be considered as a properly functioning artefact (in a sense, it doesn't function at all since it declares that everything is true, and thus also declares that everything is false). Of course, when all theories that are negation inconsistent are also absolutely inconsistent, the foregoing automatically also applies to theories that merely contradict themselves. This is no longer so in a paraconsistent logic, and in that case a different argument against theories that are negation inconsistent is needed.

The traditional view is that contradictions cannot be accepted because they are necessarily false; not just because they lead to triviality. This is the Tarskian orthodoxy (Tarski, 1944). Dialetheists agree that all contradictions are false, but also claim that some of these contradictions are true as well. Floridi follows the

[42] See e.g. Quine (1960) on the maxim of translation.

[43] The reasoning goes as follows: assume that both "*p*" and "not-*p*" are true. Because "*p*" is true, "*p* or *q*" is also true. Yet, because both "not-*p*" and "*p* or *q*" are true, it is "*q*" that must be true. Since "*q*" could be any formula, every contradiction entails any formula.

Tarskian diagnosis, but additionally argues that even if a demiurge could create an inconsistent artefact, the user of that artefact wouldn't experience the artefact as a correctly (though inconsistently) functioning artefact, but as a plainly malfunctioning artefact. When we design a model of the world, we want to use that model as an interface for successful interaction with the world. Yet, an inconsistently functioning artefact is antithetic to such successful interaction, and therefore doesn't function correctly (Floridi, 2010d).

The disagreement between Floridi and the dialetheist is interesting in its own right, for the dialethist can point to a hidden assumption in how Floridi understands correct functioning: an artefact functions correctly whenever it (a) does all that one expects it to do, and (b) refrains from doing what one doesn't expect it to do. In Floridi's argument, it is clause (b) that does the work, but the dialetheist will almost certainly reject this clause.

We can illustrate how this disagreement bears on a difference between conceptual engineering and conceptual analysis by looking at a few well-known conceptual paradoxes. Let's start with the barber, a paradox formulated by Bertrand Russell. The paradox concerns the barber of Tombstone who, according to the story, shaves all and only those men in Tombstone who do not shave themselves. Assuming that this barber is himself a man who lives in Tombstone, we can ask who shaves the barber. Since he either shaves himself or doesn't shave himself, we need to consider two cases. If we assume that he shaves himself, then by the above description of who is and isn't shaved by the barber, we have to conclude that he doesn't shave himself after all. If we assume that he doesn't shave himself, then by the same description we have to conclude that he does shave himself. Conclusion: the barber shaves himself if and only if he doesn't shave himself, but since he either does or doesn't shave himself he both does and doesn't shave himself.

The standard diagnosis of this paradox is that there cannot be such a barber. A contradiction was reached because we started out with a description of something (or someone) that cannot exist. In Floridi's terminology, this barber is an artefact we cannot construct.

Our next paradox is the liar. A liar sentence is a sentence that claims its own falsity, and which, given an intuitively plausible assumption about truth, allows us to derive a contradiction. We start from two basic principles, one about liar sentences, the other about the truth-predicate:

(1) A sentence L is a liar-sentence if and only if it is logically equivalent to not-True "L"

(2) "True" is a truth-predicate if and only if (for every A) A is logically equivalent to True "A".

As we did for the barber, we can then ask whether the liar-sentence is true or false. We start with the hypothesis that "L" is true, and follow a similar pattern:

(3) True "L" (hypothesis)

(4) L (by principle 2)

(5) not-True "L" (by principle 1)

Next, we start from the hypothesis that "L" isn't true, and reason analogously:

(6) not-True "L" (hypothesis)

(7) L (by principle 1)

(8) True "L" (by principle 2)

Jointly, these two simple pieces of reasoning show that the liar-sentence is true if and only if it is not true:

(9) True "L" if and only if not-True "L" (by 3-5 and 6-8)

Yet, by classical logic, this is just the same as saying that "L" is both true and false:

(10) True "L" and not-True "L" (from (9))

The traditional diagnosis of this paradox is in part analogous to that of the barber. One of the basic principles we started from must be false, but in contrast to how we evaluated the barber, we do not say that liar-sentences do not exist, but rather reject the existence of a truth-predicate that complies with the following general principle or truth-schema:

(Truth) "*p*" is True if and only if *p*

From the perspective of classical logic it is obvious that there cannot be such a truth-predicate, for including it in our language would lead to triviality. What is more interesting is that from the perspective of a paraconsistent logic there is no need to stick to the same diagnosis for both paradoxes. Even if the barber of Tombstone cannot actually exist, there are fewer good reasons to assume that we cannot have a non-trivial truth-predicate with paradoxical properties. This indicates an important point of disagreement between the orthodox view Floridi adheres to and the position adopted by many dialetheists. According to Floridi, semantic artefacts should meet the same logical requirements as real (physical) artefacts, and this is enough to dismiss the semantic machinery that leads to paradox. According to the dialetheist, semantic artefacts with inconsistent properties are perfectly legitimate because they capture essential features of our informal understanding of truth. It is therefore sensible to accept some of the unintended consequences – Beall (2008) calls them spandrels – to retain the full strength of our naive notion of truth. Here, the difference between conceptual engineering and analysis comes to the surface: prototypical arguments for dialetheism are grounded in conceptual analysis, and therefore accord much importance to pre-existing concepts. From a constructionist perspective there is no need to preserve such concepts, and therefore less pressure to embrace inconsistency.

12.3c Levels of abstraction

When logic is understood as the study and codification of correct inference, it is tempting to think that logic aims to separate the arguments that are really valid from those that aren't. That is, logic is meant to provide the final word on which arguments are good and which are bad. Such descriptions are misleading, as they ignore the fact that because logicians use formal techniques to determine which arguments in a formal language are valid, their results do not obviously relate to arguments that are formulated in natural language. When the gap between a formal model and the natural language phenomenon it is a model of is properly acknowledged, it becomes much harder to think of (formal) logic as the final arbiter on issues surrounding correct inference: Logic then becomes a modelling tool as is any other formal method employed in the sciences. This "logic-as-modelling" view was most explicitly defended by Stewart Shapiro. He claims that:

> … with mathematical models generally, there is typically no question of 'getting it exactly right'. For a given purpose, there may be bad models–models that are clearly incorrect–and there may be good models, but it is unlikely that one can speak of the one and only correct model. There is almost always a gap between a model and what it is a model of.
> (S. Shapiro, 2006)

This insight allows us to cross the line between the use of logical methods and the method of levels of abstraction.

The primary methodological maxim of the method of levels of abstraction is that philosophical questions cannot be answered when the LoA at which they are formulated (and at which the answer should be given) isn't made explicit:

> (Levels) Always make the relevant level of abstraction explicit.

In the prelude we have seen that we can define different kinds of formal languages. In a first-order language, the alphabet or non-logical part of the language will consist of individual constants, individual variables, and predicates (and relations). The atomic expressions that can be formulated on that basis are much like the typed variables used in the method of the levels of abstraction; they are the basic entities that can acquire a meaning. Given an interpretation for that language, which is a way of assigning individual constants to members of a domain (the objects we can talk about) and predicates to their extension in that domain (the objects that satisfy the predicate), the atomic expressions can acquire a meaning. As such, interpreted atomic expressions are much like the observables used in the method of levels of abstraction; they are the basic entities that have meaning.

Take for instance the plane example from Chapter 2. For every relevant feature of the plane, we can introduce a corresponding predicate e.g. "P" for expensive, "I" for internal condition, or "E" for external condition, etc. Similarly, the level of abstraction for each agent could then be modelled as the non-logical part of the language that the agent in question uses to describe the plane they are talking about.

Given this natural correspondence between the basic building blocks of the method of abstraction and the atomic expressions of first-order languages, formal languages (and by extension formal consequence relations) are an excellent tool for making the LoA at which one operates precise.

12.4 An informational perspective on logic

The development of an informational perspective on logic can be framed as an answer to the following question:

> (Q) To what extent do the truth-based and the information-based criterion we use to evaluate arguments agree?

Despite the traditional view that "not going beyond one's premises" and "not moving from true premises to false conclusions" are really the same criterion, certain principles of classical logic are a good reason to doubt this type of identification. Consider the following arguments:

(1) If *q*, then (*p* or not-*p*). Or, if Alice smiles Bob blushes or Bob does not blush.

(2) If (*p* and not-*p*), then *q*. Or, if Bob does and does not blush, Alice smiles.

(3) If *p*, then (if *q*, then *p*). If Alice smiles, then if Bob blushes Alice smiles.

These are known as paradoxes of material and strict implication or implicational paradoxes (Read, 1995), and expose several limitations of classical logic.

The first one is valid because the conclusion is always true; whether or not Alice smiles, Bob will always either blush or not blush. The second is true because its sole premise is contradictory and hence can never be true; Bob can never both blush and not blush, so it doesn't matter whether or not Alice smiles. The third one is true because the truth of "*p*" guarantees its own truth, and hence also the truth of the implication "if *q* then *p*". If Alice is already smiling, it is still true that she smiles if we additionally assume that Bob blushes.

As such, each of these principles complies with the requirement that valid arguments can never have true premises and a false conclusion: (1) always has a true conclusion, (2) always has a false premise, and (3) can never have a false conclusion without also having a false premise.

Yet, these principles have been called paradoxes because in each case there is no connection whatsoever between *p* and *q*. Indeed, *q* is entirely arbitrary, and this clashes with the initial suggestion that, for instance, the content of "*p* or not-*p*" should be contained in the content of "*q*". This is not a knock-down argument against the validity of the principles (1)–(3), but it does reveal that the happy coincidence of truth-preservation and the inclusion of conclusions in premises may just be a peculiar feature of classical logic. The reason is that, according to how classical logic functions, the content of a logical truth like "*p* or not-*p*" is null, and therefore included in any other content; and that the content of a logical falsehood like "*p* and not-*p*" is maximal, and therefore includes any other content. If one already accepts classical logic, one might not be troubled by this diagnosis, but if one finds the paradoxes of material and strict implication disturbing, the classical analysis of the content of logical truths and falsities isn't reassuring either. We make these insights more precise by taking a closer look at the metaphor of logical space.

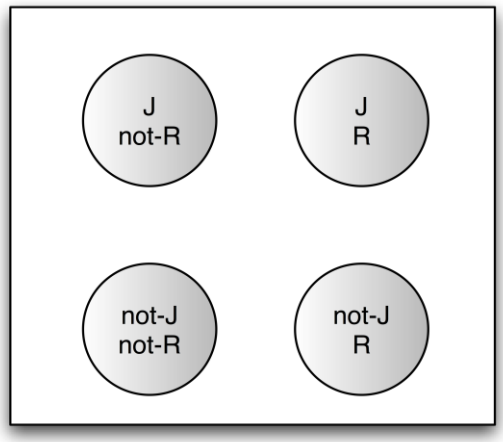

Figure 10: Logical space with four possibilities

Consider once more our example where Alice promised Bob a reward if he completed the job. Before Alice made this promise, 4 situations were possible: Bob did the job, but didn't get a reward; Bob did the job and did get a reward; Bob didn't do the job, and didn't get a reward; and Bob didn't do the job, but did get a reward. We can depict this as follows in Figure 10.

Once Alice announced that Bob would get a reward upon completion of his job, the upper-left situation—job, but no reward—is no longer a possibility. This is depicted in Figure 11.

Diagrams like those in Figure 10 and 11 can be used to reason about what does and does not follow from Alice's promise: We can "see" that once the upper-right possibility is removed:

- the only possibility where Bob does the job is also a possibility where he gets a reward,
- the only possibility where Bob doesn't get a reward is also a possibility where he didn't do the job, and

- there is still a possibility of getting a reward while not doing the job.

It is easy to see that these three points nicely match the good and bad arguments with which we started out this chapter (see also the exercises at the end of this chapter).

In addition to being a tool to evaluate arguments, the same diagrams also provide us with a nice model to think about information. As was already explained in one of the introductory chapters, Dretske proposed to extend one of Shannon's main insights and claimed that the informativeness of a piece of information decreases with the probability of it being true. When we use diagrams to depict the possibilities that are available in a given logical space, we rely on an analogous, but non-probabilistic relation (we call such approaches qualitative, as opposed to the quantitative approach of probability theory) between informativeness and likeliness. This relation is summarized by the Inverse Relationship Principle (e.g. (Barwise, 1997)):

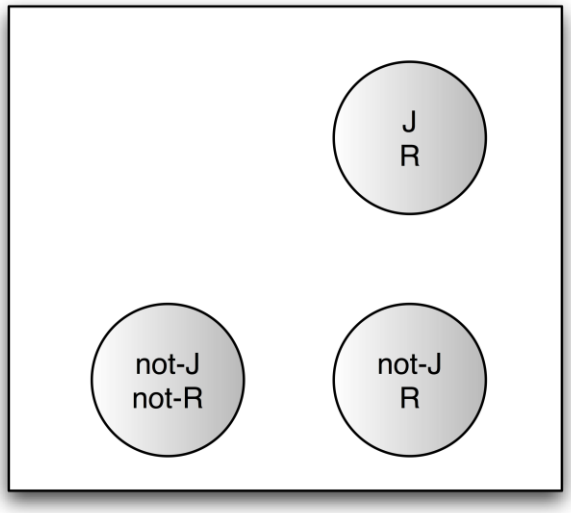

Figure 11: Logical space with three possibilities

(Inverse Relationship Principle) Whenever there is an increase in available information there is a corresponding decrease in possibilities, and vice versa.

Given this principle, we can identify the content of *A* or the information in *A*, with a set of possibilities, and note that more information means fewer possibilities. Let us now take a closer look at what possibilities are, and what the logical space metaphor amounts to.

Possibilities and sets of possibilities play a central role in many theories of meaning. The notion of a proposition or propositional content is a close relative of the notion of semantic information. We often say that the information conveyed by *A* is just the proposition expressed by *A*. To focus on the propositional content of *A* is to ignore certain syntactical specificities of how *A* is formulated. For instance, we do not care whether *A* is in fact "Alice smiles and Bob blushes" or "Bob blushes and Alice smiles"; both express the same proposition. One highly influential way of looking at propositions is as sets of possibilities or proportions of a logical space (Stalnaker, 1984). On that account, the proposition expressed by *A* is just the set of cases in the logical space where *A* is true. As we have seen in the technical prelude, cases are just things where formulae can be true (or false), and a logical space is just the set of all acceptable cases. What counts as an acceptable case is something that is open to debate. The most popular account takes cases to be possible worlds: cases that, just like the complete and consistent cases we introduced in the technical prelude, decide every issue.

When we use sets of possibilities to represent the content of pieces of information, this has certain non-trivial consequences. One such consequence is that if two formulae *A* and *B* are true in exactly the same set of possibilities, they express the same proposition. This will also mean that if fewer cases are acceptable, we will be able to distinguish fewer formulae, for if we want to distinguish *A* from *B*, we will need a case where one is true and the other one false.

When possible worlds[44] are the only acceptable cases, we do not only miss out on many potentially interesting distinctions, but we also end up with the implicational paradoxes. Unsurprisingly, if we are less restrictive in our inclusion of cases, we cannot only deal with the implicational paradoxes, but we can also say something about how logic is related to information-extraction.

Given a logical space, the information conveyed by *A* can be understood in two ways: (1) as the proposition expressed by *A*, or (2) as the relative complement of that proposition in the whole logical space (the possibilities excluded by A; see Figure 12). Either way, the result is a proportion of the total logical space. On the former account, a smaller proportion means more information; on the second account, a larger proportion means more information.

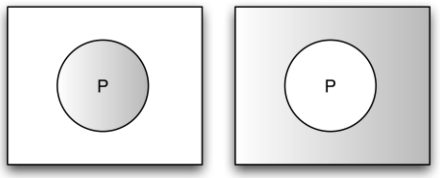

Figure 12: Proposition expressed by *A* (left) and possibilities excluded by *A* (right)

Consider again the implicational paradoxes. If cases consistently decide every issue, a logical truth will be true all over the logical space, and a logical falsity will be false all over the logical space. Since the complement of the total space is the null-space and vice-versa, the information in logical truths and logical falsities are, respectively, the null-space and the total space. Logical truths do not exclude anything; logical falsities exclude everything. These features are associated with classical logic, but also with how we think about logical spaces. As long as logical truths are all-over true, and logical falsities are all-over false, similar features will arise. We do not only have "If *q*, then (*p* or not-*p*)" because "*p* or not-*p*" is true anyway, but also because it is entirely uninformative and thus no more informative than any *q*. Analogously, we do not only have "If (*p* and not-*p*) then *q*" because "*p* and not-*p*" is false anyway, but also because it is maximally informative and therefore more informative than any *q*. So presented, the implicational paradoxes can also be understood as informational paradoxes. Principles (1) and (2) are related to, respectively, the "scandal of deduction" (logical truths cannot be informative) and the Bar-Hillel Carnap Paradox (contradictions are maximally informative).[45]

A standard way out is based on the dilution of the logical space: by including cases where logical truths can fail, and often also cases where logical falsities can obtain (these can, but need not be, the same cases), the above reasoning is blocked. If, moreover, one stipulates that logical truths need not be true across the whole space, but only need to be true in a subset of that space (the "logical" cases), a logical space can even be diluted without a change in logic. Talking about information might require some cases that are not required for logical theorising, and the dilution of logical space is the most straightforward way of avoiding the conclusion that logical truths and logical falsities convey, respectively, minimal and maximal informational content. Of course, if one also wants to avoid the implicational paradoxes, it is best to adopt a uniform solution for informational and implicational paradoxes.

An additional virtue of the dilution of a logical space is related to our ability to distinguish between the content of different pieces of information.

The problem with a classical or possible-worlds account of information and consequence, one might say, isn't so much that logical falsities are replete with information or that logical truths are devoid of information, but

[44] Or some other type of case that consistently decides every issue, like for instance the state-descriptions used by Carnap and Bar-Hillel (1952).

[45] Floridi (2004c) proposes a solution to the Bar-Hillel Carnap paradox that is based on the veridicality of information (see Chapter 7): since contradictions cannot be true, they cannot be informative either. Needless to say that this solution also requires us to give up the simple logical space metaphor we described in this section.

rather that in either case they do not merely convey the same amount of information, but can be regarded as the same piece of information. As far as the classical theory goes, there is just one contradictory content, and just one logical content. Again, this property of our theory of informational content is the pendant of a well-known feature of classical logic: logical truths are logically equivalent, and so are logical falsities. The problem, in both cases, is a failure to discriminate between logical contents, or between counter-logical contents. Even if one agrees that, come what may, logical truths are true, and logical falsities are false, we often want to resist the conclusion that, for instance, two inconsistent theories like the infinitesimal calculus of Leibniz and Newton (Brown & Priest, 2004; Norton, 1987) and the old quantum-theory have the same informational content. This redirects the initial worry about the null content of logical truths and the maximal content of contradictions, to a more general worry about what can be told apart in terms of informational contents. Because from a purely formal perspective the inability to discriminate between certain contents is just a different guise of the initial paradoxical features of classical logic and the classical theory of informational content, the dilution of logical space solves both problems at once. Still, by focusing on the distinctions that become available by diluting the logical space, we can come up with a more balanced diagnosis of the use of possible worlds. It isn't necessarily mistaken (just as classical logic isn't necessarily mistaken), but it is surely inadequate for certain purposes, namely when we want to tell certain logical truths (or certain logical falsities) apart. This opens the door for a pluralist account of informational content, but also suggests an attractive picture of information extraction. We have seen that our ability to make distinctions is tied to the availability of the right kind of possibilities in our logical space. Highly simplified, many possibilities lead to many distinctions. At the same time, having many possibilities also makes it harder for arguments to be good.

This type of trade-off is well-known, and arises for many types of formal modelling. If we make abstraction of more features of the system under investigation, it becomes easier to arrive at conclusions, but our conclusions will be less precise or discerning. Conversely, if we make fewer abstractions, our conclusions are potentially more discerning, but it is also harder to reach these conclusions. When deductive inference is made precise by making clear the LoA at which we operate (settling on an interpreted language, and thus also on a consequence relation), the situation is exactly the same as for any other modelling enterprise: We simply can't have it all. When we evaluate a logical system, we first need to balance the opposite virtues of logical discrimination and deductive strength, and then decide which logical system is the most appropriate for a given purpose.

12.5 Exercises

1. Take a new look at the three paradoxes of material implication, and try for each of them to find, first, an interpretation for *p* and *q* that makes the principle intuitively false, and, second, another interpretation that makes the principle intuitively true.

2. We already gave an example of inconsistent theories that we might want to keep apart. Can you also think of an example of logically true theories that we also might want to keep apart? Why would that be useful?

12.6 Suggestions for the exercises

1. Use the diagrams from the Figures to explain why *modus ponens* and *modus tollens* are good arguments, but *denying the antecedent* is a fallacy.

2. Use *q*s that are utterly absurd. Experiment with *p*s and *q*s that are more or less related. Consider mathematical theories, or other theories that are true of necessity because they rely on proofs.

12.7 Further reading

Good introductory reading includes Priest (2008) and Read (1995). Nice advanced pieces include Van Benthem and Martinez (2008) and Allo and Mares (2012).

Chapter

13

13. COMPUTATION

Theoretical and practical information processing

13.1 Introduction: how are information and computation related?

> *'It would appear that we have reached the limits of what it is possible to achieve with computer technology, although one should be careful with such statements, as they tend to sound pretty silly in 5 years.' John von Neumann, ca. 1949.*

To many computer scientists as well as computer users the answer to the question of how information and computation are related may seem trivial: *computers traffic in information*. But to get to the bottom of this question, we first need to understand better what we mean by computation and what we mean by information. In some very loose sense, it is true that computers traffic in information: they are built to store, manipulate and exchange information. But when we dig deeper, this qualification may seem insufficient, if not incorrect. Consider the situation where Alice is checking on the internet for the shortest way to travel to meet her friend. Connected to the Railway Company Website, Alice is entering information: dates and hours, places, and discount options and she is expecting to obtain departure and arrival hours and prices. On the other side of her laptop is Bob, the server machine of the Railway Company, answering Alice's enquiries, matching them against an optimal solution. What Bob is receiving and processing are really elements of certain lists of points in a graph; numbers associated by certain operations to some subsets of those lists; and subsets of those numbers when additional conditions are provided. For Bob these hardly are, respectively, locations, prices and discounts. In other words, what is the nature of the entries by Alice, and how are they related to the elements treated by Bob? This question can promptly be better qualified: how is the *information* entered (and required) by Alice related to the *data* processed by Bob? What *kind* of information is ultimately the object of computational processes?

From this first example, it seems clear that "information" should be distinguished from "data", although they are intimately related. In computer science, "raw" data typically refer to any variable or signal that is not an instruction. A *computer program* can be viewed as *a mechanism that consumes, processes and produces data.* One task of the present chapter is to clarify the fundamental operations of computing processes as data manipulators and to explain which theoretical and practical principles are at the core of a computer program. To better highlight the relation between data, instructions and programs, the next section provides some background on the conflation of information and digital computation. We then turn to discuss the theoretical foundations of computability, a research field in its own right, which has at its very

core the formal treatment of the question: *what does it mean to compute*? The answer to this question is directly relevant for understanding today's information technologies, and for explaining this we shall introduce the basic concept of a digital computing machine developed by Alan Turing that is now known as a *Turing machine*. (A more basic introduction is also given in Chapter 1.)

The second task of this chapter is to identify the (still) missing link between data and information i.e. the connection between quantifiable data and their meanings or, if you prefer, the explanation of how numbers and lists crunched by Bob become places, prices and times for Alice. This task has been a major problem for theories of semantic information. In Shannon's information theory, data simply means signals or messages (Floridi, 2008, p. 119). In the context of conventional digital computers, data may be symbols or numerals, but also pictures and sounds. But, strictly speaking, the same "raw" data (say, a string) may be acted upon as both input data and an instruction (i.e. a part of the program) in a single run of some program. That is, some string (e.g. "exit") may be entered as data by the user (and its value may be stored by some variable, e.g. var1 = "exit") and subsequently that string can be processed by that program as an instruction. The meaning of the particular string seems to be context-dependent.

From Alice's perspective, informativeness corresponds to the data being precise, correct and timely; in other words, to be the information she *expects*. But from Bob's measuring perspective, the output of its processing is dictated by instructions; hence, only *surprising* data can really be informative. Paradoxical as it may sound, contradictory timetables and a TV set emitting white noise are most informative, on some theories of information! In Section 13.4, we offer an overview of some approaches explaining what turns data into information and spell out how digital computation may be construed as the processing of information.

However, there is no reason in principle to restrict *computation* just to *digital* computation, though it is certainly the most common type of computation used today. While the traditional dichotomy is between digital and analogue computation, other sorts include quantum computation (i.e. computation that makes use of quantum mechanical phenomena), neural computation (i.e. computation of whichever kind that happens in the brain), DNA computation (i.e. computation that takes advantage of DNA molecules for its processing) and natural computation (i.e. computation that employs natural materials for its processing). The remaining types may or may not fall under one of the two broad categories of digital and analogue computation. Our focus in this chapter is mainly on digital computation, though some attention is also given to analogue computation. Whilst the objects of computation constitute a fascinating topic on their own, for simplicity we shall assume that the traditional digital/analogue dichotomy suffices for classifying computing systems.[46] In Section 13.5 we discuss the notion of analogue computation and explain how it differs from its digital counterpart. The subsequent section traces the relation between computing processes and artificial intelligence, explaining how "intelligent" machines use information (while *the Terminator* is not discussed here, game winners, such as *Deep Blue* and *Watson*, are).

The main aim of this chapter is to illustrate the strong conceptual and practical relation between information and computation. We shall see that any computing process, from those executed by our digital devices (such as PCs, tablets and phones), to those executed by analogue devices (such as traditional watches and slide rules), can be understood and explained in terms of the data and information they process and the output they give us, in turn providing a new view of their nature.

46 For some further discussion on this topic see (Copeland, 1996; MacLennan, 2004; Piccinini, 2008a).

13.2 A brief background on the conflation of computation and information

The conflation of information and computation may be traced back at least as far as the early twentieth century. In *Cybernetics* (see Chapters 1 and 10), self-governance and control in mechanical and biological systems were explained by studying how they process and reuse information.[47] On the one hand, this called for a mathematical definition of the object of such processes, which was offered by Shannon's information theory (introduced in Chapter 1). On the other hand, it also required a precise understanding of the *ways* information can be processed, which was in turn offered by classical computability theory, whose forefathers include Alan Turing, Alonzo Church, Kurt Gödel and Emil Post. The motivation for conflating computability theory and information theory was the suggestion that the brain's information-processing capacities and its limitations could be measured quantitatively and not just compared qualitatively. The result of this venture was a fusion of information-theoretical language, notions of control and feedback from contemporary cybernetics, as well as notions (such as program, simulation and formal language) from computability theory.

In what is presently known as cognitive science, ideas from information theory needed to be supplemented by ideas from computability theory. Experimental psychologists and some neurophysiologists in the mid-twentieth century adopted information-theoretical language, since it offered the right terminology (such as coding and transformation), tools and quantitative measurements. But since information theory in isolation was insufficient to explain mental processes in terms of internal representations, it had to be supplemented somehow (Boden, 2008, p. 744). This gap was bridged by importing computation-theoretic language into the cognitive scientific discourse. Computability theory offered two main advantages in this venture. Firstly, it seemed to provide the right mathematical tools for modeling neural activity in the brain. Secondly, it had the "right" language to describe cognition: programs, implementation, formal languages, logical formalisms and so on. These were the early foundations of cognitive science, as we know it today.

We have already reviewed Shannon's information theory in Chapter 1, and the formation of cognitive science is described in more detail in section 10.2. Let us now turn to survey the fundamentals of computability theory and the notion of a Turing machine.

13.3a Computability: the theoretical basis of computation

Computation and even more specifically, digital computation is by no means an unequivocal concept. There are many subtleties that require further discussion, but we only touch on a few of them here. To start with, we can, and should, distinguish computability from physical computation. Computability is an abstract concept and was the main focus of mathematicians such as Turing, Church, Post and Gödel in the 1930s. Physical computation, on the other hand, is the physical manifestation of a computing process in real-world systems. Clear examples are a conventional digital computer, an iPad, an Android smartphone or the system run by Bob that Alice is using to book her train travel.

An informal and intuitive understanding of a computing *process* is described as follows. A computing process is a procedure executed in a finite number of steps, which at each step presents a finite and

47 For an easy-to-digest introduction to cybernetics see (Cordeschi, 2008).

complete set of rules to be applied for any possible input, which allows no random choices in the execution of rules to establish its next step. In our simple example above, when Alice enters the departure and arrival stations in the Railway Company webform (i.e. the input data), Bob selects the subset of train connections corresponding to the requested departure-arrival selection (this operation only, rather than some random selection inverting departure and arrival stations, unless prompted otherwise) and outputs the relevant timetable (as the final step).

Of course, a computing process can also be defined more formally, such as a computing process is an algorithm that instantiates a function (operation) over numerals. But even this definition is far from being conclusive. Other examples that explain precisely the same concept are simple computing with an abacus, the highly formal lambda-calculus by Church, and the fascinating idea of a machine performing *any well-defined routine of operations* corresponding to numerical functions.

To understand why there are many *equivalent* notions of computations, we need to explain in what sense a function constitutes a formal specification of a computing method. Think of a set of numbers, for example, the set of all natural numbers, or any subset thereof, like the set of odd numbers or the set of numbers corresponding to days Alice wants to take a train. A computing method would be a process that lists *all* and *only* those elements in any such set (such as Bob's selection of all and only trains departing on the days chosen by Alice). In fact, a function is a mathematical object $f(n)=m$, which takes positive integers as arguments (n) and gives members of a certain given set as values m.

For example, the successor function $f(n)=n+1$ lists the set of all natural numbers. It is the function that takes any number n as argument and returns the successive value. It is easy to see how f is the function that returns the whole set of natural numbers when $n=0$. For given 0 f returns 1, given 1 f returns 2, and so on. We also say that such a function *codes* an arrangement of the members of the set (of the natural numbers). In turn, any set whose members can be listed (i.e. are enumerable) by a definite method, is the result of applying some function to a set of positive integers. This shows that functions of positive integers represent a method to generate the content of any given well-defined set of numbers, by providing the set of rules, or instructions, to construct it.

This notion of defining and executing instructions corresponds to the idea of *effective computability: a function f is called effectively computable if a list of definite and explicit instructions can be given to compute the value of f(n), for any argument n in the range of positive integers*. Hence, at each step and for each argument, there is an explicit formulation of instructions to be followed in order to obtain a result. External problems, such as time and energy, are irrelevant. One way to translate this is by saying that the entire process could be performed *mechanically,* without any human insight. If there is a function to extract the timetable of trains on the days Alice wants to travel, then Bob, the server, should be able to do it. This is what Turing was able to define so precisely and eloquently.

13.3b Turing machines

As already mentioned in Chapter 1, Turing is considered by many to be the father of computer science and his work is crucial for understanding the relation between computation and information.[48] While Turing's focus was computable numbers as well as mechanical procedures and not the concept of information per se, his work

[48] The roots of the first computing system may be traced back to Gottfried Leibniz's calculating machine (and perhaps even earlier than that). To consider Turing *the* father of computer science is undoubtedly a gross simplification, but for our purposes here it is convenient to start the discussion with Turing.

reflects the intimate relation that exists among them. This relation is made particularly conspicuous in introducing the concept of universality, which unifies data and computer programs by way of the universal Turing machine. An important take-home message from this present section is the generality of Turing's invention: if a specific problem can be solved by an algorithm[49], then there exists a Turing machine that can solve it. But before discussing the concept of the universal Turing machine, let us begin with the problem that Turing was tackling and how it led him to inventing these machines.

Turing's work in the mid-'30s was inspired, as was Church's, by the decision problem, which had been formulated by the mathematician David Hilbert in 1928. The decision problem could be formulated as follows. Does there exist an effectively computable procedure which, applied to any assertion of first order logic, could decide whether it is true? Both Church (1936) and Turing (1936) published papers independently, showing that such a procedure is impossible, and hence verifying the undecidability property of formal systems proved by Gödel. For his result, Turing came up with the automatic machine or a-machine (which we shall now refer to as a Turing machine) that mimics a human who is calculating using just pen and paper. This was the precise mechanical understanding of effectively computable procedure.

According to Turing, the operations of the human computor (as distinct from an artificial computer) may be completely mechanised by breaking the rules of computation into a series of basic sub-rules. The human computor was seen to facilitate her calculations by using a notebook and focusing her attention at any given moment on a particular page. By following the instructions she may alter that page or turn to another page. The alphabet of symbols available to the human computor may be assumed to be finite and the content of each page can be replaced, in principle, by a single symbol. By envisaging a notebook large enough so that the human computor never reaches the last page, the result is an infinite running paper tape. A computation would be understood as a finite sequence of operations on symbols.

The Turing machine was an abstract, idealised mechanical device representing a human computor, equipped with unlimited storage capacity (an infinite tape to write on), whose operations are determined by discrete, effective steps so that at each step it is entirely defined what the computor is allowed to do, and a limited set of possible actions (defined by its table of instructions). It was, perhaps surprisingly, proven that such a general and simple machine would be computationally equivalent to almost any conceivable digital computing system. Turing machines perform calculations on the tape, by way of a head (or a scanner), a state "register" and a table of instructions (Turing, 1936, p. 231). The tape is divided into squares, each of which may either be blank or bear a symbol. The head goes through the squares of the tape, scans a symbol, possibly erases that symbol and writes a new one and then moves one square either to the left or the right. The state "register" is the machine's equivalent of the human computor's "state of mind", which is assumed to be in one of a finite number of states (Turing, 1936, p. 250). This "register" may be conceived as being part of the machine's head. The machine's table of instructions contains the state-transition rules or instructions. The first two columns of the table are the m-configuration and the scanned symbol. The m-configuration is a finite number of conditions, which the machine is capable of. The pair (m-configuration, scanned-symbol) is called the machine's configuration and uniquely determines its behaviour. The last two columns are the operation the machine executes when it is in a particular configuration and the final m-configuration the machine enters upon completing its current operation (Turing, 1936, p. 234). In the next section we return to the distinction between data, broadly construed, and semantic content, and will show how the table of instructions becomes crucial.

[49] This is a notion that predates Turing's work and can be roughly described as a "recipe" for performing a function-based calculation that can be followed "mechanically".

To explain further how the Turing machine works, let us use an example of a machine that computes the successor function on some integer number *n*. The machine table below (figure 13) describes four transition rules that suffice for this computation, with some further assumptions about how the machine operates. For simplicity, this Turing machine may only have one number represented on the tape. The machine starts with its head scanning the leftmost "1" symbol of a sequence of a block of *n* 1s. The block of 1s representing the argument of the function is delimited by an occurrence of the symbol "0" at the end. The number zero is represented by a single "1", the number one is represented as "11" and so forth: the sequence 01110 codes the input 2. Only the symbols "0" and "1" may appear on the tape. The machine's head is positioned at the leftmost "1" symbol of the sequence when it halts. As for the other conventions used, state-1 is the initial state of the machine and state-3 is its terminating state. "1" in the operation column means erase the scanned symbol and replace it with a "1". "R" in the operation column means move one square to the right and similarly for "L".[50] The Turing machine described in the table below is a special-purpose machine, that is, a machine that executes some specific, well-defined function.

m-configuration	Scanned symbol	Operation	Final m-configuration
state-1	0	1	state-2
state-1	1	R	state-1
state-2	0	R	state-3
state-2	1	L	state-2

Figure 13: An instruction table for a Turing machine that computes $f(n)=n+1$

Besides the specific technical details of how a Turing machine works, Turing was able to show that effective computation can be formulated using an arbitrary machine that is subject to some finiteness restrictions: a finite number of states and a finite number of possible symbols on the machine's tape. It is the generality of the Turing machine, rather than its particular design, that was such a crucial step in the evolution of theoretical computer science. A Turing machine can simulate any classical computing device, though perhaps not as efficiently in terms of its runtime. The thesis that effective computability is entirely satisfied by the mechanical procedure, as envisaged by Turing, is known as the Turing Thesis. It states that any effectively computable function is computable by a Turing machine. It is justified, on the one hand, by the fact that for any computable function, a procedure by a Turing machine can be defined. On the other hand, it is justified by the provable equivalence of mechanical, that is, effective, computation with other notions of computability, such as lambda-definability (an equivalence that generates the so-called Church-Turing Thesis). One of the most interesting results in computer science is that Turing machines, lambda-calculus, Post's production systems and other formalisms of computability are all extensionally equivalent in the sense that they all identify the same class of functions as computable.

The relation between digital computation and information becomes even more obvious when considering universal Turing machines. The universal Turing machine exhibits another extremely important form of generality. Turing showed that it is possible to construct a single universal machine U that can be used to compute any function that is computable on a special-purpose machine M (Turing, 1936, pp. 241-242). If U is supplied with a tape on the beginning of which is the machine table of M, then U will compute the same function as M. The exact construction of U exceeds the scope of this chapter.[51] Suffice it to say that the

[50] Several Turing machine simulators may be found on the Internet. Try one at http://ironphoenix.org/tril/tm/
[51] The interested reader can find references for further reading at the end of the chapter.

operation of the universal Turing machine is explicable in terms of its execution of the instructions of some special-purpose Turing machine. While the latter may be conceived as an algorithm (i.e. an effective procedure) or a conventional computer program[52] that was designed for a particular or special-purpose task, the universal Turing machine may be conceived as a general-purpose digital computer that executes programs (i.e., descriptions of other special-purpose machines). Unlike the special-purpose Turing machine, the universal machine is conceived as a soft-programmable system, which can simply be reprogrammed by erasing its existing program, of some special-purpose machine M1, from its tape and inscribing another program, of a different special-purpose machine M2, on its tape.

For the purpose of our discussion of information and computation, an important consequence of the design of the universal Turing machine is the resulting distinction between "raw" data and programs, or sets of instructions. Turing's concept of universality unifies "raw" data and programs. On the one hand, U takes M's table of instructions (encoded on U's tape) as a program to be executed. On the other hand, U takes the input data of M (also encoded on U's tape) as the "raw" data, which are operated on by the program. Both the "raw" data and the program are encoded in a similar manner on U's tape and the essential distinction between them is the functional role they play in the operation of U. In the following section, we elaborate on the type of information the program qualifies as.

13.4 Digital computation as information processing

Computation is often said to be a specific kind of information processing. But which notion of information processing can be used to analyse computation, and in particular digital computation, is controversial. This problem crucially depends on what we take information and its processing to be. This is yet another reminder of the crucial distinction to be made between data and information as discussed in section 13.1. It is one thing to claim that digital computation is the processing of data (that may, but need not, be structured and meaningful), and it is another to claim that it is the processing of truthful semantic content (the latter claim is harder to defend).

There are several options for interpreting information, some more problematic than others, even outside the context of digital computation. Some simply equate "information" with "data", for instance in Shannon's interpretation of "information" as a message that is selected from a set of possible messages with a probability distribution over it. The question above then reads as "is digital computation the processing of Shannon information?" Despite the applicability of Shannon information to many aspects of digital computation, it is primarily concerned with data and data encoding relative to some distribution of probabilities and it is not suitable for accurately explaining the phenomenon of digital computation, broadly construed (Fresco, forthcoming; Piccinini & Scarantino, 2011). Digital computation need not have the fundamental probabilistic characterisation defining Shannon information. In fact most digital applications we are used to, such as calculator applications, text-editors and so on, are deterministic – at least in principle.

But the real divide between "data" and "information" is the semantic interpretation of information. First, recall Alice and Bob. Alice enters some parameters concerning dates, hours, departure and arrival stations, as well as discount options. Bob, on the other hand, receives all these parameters as data, broadly construed. But a closer inspection reveals that not all the data have the same "role". Dates, hours and the relevant stations can be construed as "raw" data operated upon by Bob. But when Alice selects: "find optimal route between departure

[52] An algorithm and a program are not the same thing. Roughly speaking, a program is an implementation of an algorithm in some specific programming language, such as C or Java.

and arrival stations" she instructs Bob to perform one computation. When Alice selects "calculate fares for the route found" she instructs Bob to perform yet another computation. If we equate semantic content with semantic information, then it is simply structured data that is meaningful for some potential agent or system (say, Bob). Semantic content is sufficient for what Floridi calls instructional information (Floridi, 2011c). It may be described as the directions to make something happen or accomplish some task, such as a recipe for making pizza. Some go one step further and insist that only truthful semantic content (what Floridi calls factual information) count as semantic information (see Chapter 7). Factual information represents facts or states of affairs (Floridi, 2007a).

In the context of examining the plausibility of instructional information as a candidate for explaining (nontrivial) digital computation, the universal Turing machine comes in very handy. A universal machine, U, can process instructional information, by means of encoding the machine table of some other special-purpose Turing machine, T, on its tape as well as some other "raw" data as the input to T. In simulating T, U reads data from the tape and simulates T's instructions on the "raw" data, which were T's input. These instructions are processed in the course of U simulating T: they are read, modified and even erased.

On the other hand, factual information as a candidate for explaining digital computation is problematic at best. In addition to the underlying data being non-empty, well-formed and meaningful (that is, the general definition of semantic information), they are also truthful. The processing of factual information by a digital computer, in the sense that is inherent to the computation performed, requires that the computer be sensitive to the truth of the semantic content processed. Some have argued that this is exactly what digital computers are good at (Newell, 1980; Pylyshyn, 1984), but others have argued against it (Piccinini & Scarantino, 2011). Whilst it is arguably implausible that digital computers inherently process factual (and hence true) information, they may still be deemed processors of digital data in accordance with finite instructional information (Fresco, forthcoming). For even if digital computers cannot evaluate truth, or ensure that truth is always preserved in the course of processing factual information, they are not required to do so in processing instructional information.

13.5 Analogue computation and information

Starting from the 1930s, digital machines were constructed in the UK (such as the Colossi or the Manchester machines) and in the USA (such as the Harvard Mark I, the ENIAC and the EDVAC) as mechanical implementations of the Universal Turing Machine. These machines would use punched cards, cathode and mercury tubes or valves as ways to input, preserve and access data in the memory. These proved to be quite inefficient methods, as they required very restrictive conditions: for example, a mercury tube would typically require the data representing a number to be positioned physically high on the tube at the moment where such a number was to be read. Digital computing improved in the 1940s and 1950s, and really got going when electronics came in, providing faster and more reliable ways to store and access data digitally, up to the introduction of transistors and integrated circuits. Today, physical limits still constrain digital machines. An interesting concept is then offered by the processing of data by analogue machines.

Analogue computation can be defined in terms of the way it treats information; it is the process of operating on continuous physical signals, analogue to the actual data being computed. Such signals can be electrical (using switches, resistors, amplifiers etc.), mechanical (using slide rules, rotors, gears etc.) or hydraulic (using pipes, valves etc.). These measuring methods are used to input standard mechanical operations of summation, multiplication, logarithm etc. In the case of electric signals, these allow the representation of the variables of a problem within a certain continuous range of voltage values, which in turn can be added, subtracted or

multiplied by means of the electronic circuit. Non-linear operations (square-rooting, multiplication of two variables etc.) can also be implemented. The crucial difference with digital computing systems lies precisely in the way information is presented to (and processed by) the machine, in terms of continuous rather than discrete values. Analogue computation does not proceed with a binary translation of the data, preserving a much higher fidelity. Instead, it involves an increased complexity in treatment of operational transmission. What kind of information cannot be coded into a digital format, and thus may require an analogue treatment?

The algorithmic model implemented by Turing machines is discrete in the sense that there is a minimal granularity of the input under which the computation cannot preserve changes in the output. For example, a digital computation in this sense might tell us if a certain point is within or without a certain space, and how to change such coordinates by moving inside or outside of the given space. A continuous function is, by contrast, one that is able to preserve small changes of the input into small changes of the output, so that, for example, a function continuous over space would be one that tells us how far a point is outside of a certain space, and how such distance changes with movement. To this aim, analogue computation needs in the first place to represent information by minimal quantities (for example, with real numbers rather than integers only); and a logical treatment of the information that admits a more fine-grained representation of states (e.g. by adding a third value to 1 and 0). In this way, analogue algorithms generalize computation by allowing continuous operations on space and time, by relying on parallel operations that maximize speed and provide more exact measurement.

Analogue machines deal with information in a very different way at many levels: acquisition, encryption, manipulation, storage and distribution. At the first level, the machine acts on the information produced and transmitted by its physical parts, in a way that makes appropriate the analogy to physical parts reacting in nature according to mechanical laws. At the level of encryption for use by the machine, there is no intermediate stage of symbolic representation of data (e.g. in terms of binary coding) and the machine reads it in its full load by means of a larger range of values, including real numbers. Storage and distribution (by reproduction) are the aspects more affected by the limitations of the non-digital medium underlying analogue machines. The logic underlying such machines is in general defined by a closure of computable functions on real values with appropriate arithmetical operations. But the determination of a common computational model for analogue machines is far less evident. And although some computational identities are known (e.g. between Shannon's General Purpose Analog Computers (Shannon, 1941) – a model based on Bush's differential analyzer (Bush, 1931) – and the computable fragment of analysis), no such thing as a version of the Church-Turing Thesis for analogue devices holds.

A very interesting thesis holds that (at least) some (artificial) neural networks are analogue computers. The complication here too is that there is no univocal notion of analogue computation either (for more on this debate see for example (Fresco, 2010; Piccinini, 2008b)). Perhaps even (some) neural networks (possibly also natural ones in animals) are typical cases of hybrid computers. A hybrid computer is understood as integrating the logical and functional aspects of digital machines with devices apt to solve differential equations from the analogue counterpart. These are in general fast machines, though with a somewhat low level of precision. The tentative analogy with (natural) neural networks is then maybe just another indication of the long-standing open problem at the basis of any direction of research in AI: can intelligence be mechanically reproduced?

13.6 Intelligent machinery and information

In addition to the Turing Machine and the Turing Thesis, Turing made another contribution to the theory and practice of artificial intelligence by way of the Turing Test. It was presented first in Computing Machinery and

Intelligence, published in 1950 in Mind, and since then has become a classic in the literature. Let us briefly outline the idea of this test, also known as the "imitation game" (see also Chapter 1). A questioner is connected to two respondents, one of which is human, the other a machine programmed to reply as a human (Turing, 1950). The aim of the questioner is to find out which one of the respondents is the machine, by analysing their answers. If the machine were able to answer in a way that makes it impossible for the questioner to distinguish the machine from the human player, then, by Turing's lights, it would be appropriate to assert that the machine is behaving intelligently, though, strictly, this would not necessarily mean that the machine displays any intelligence in the human sense. The literature on the adequacy of the test as a satisfactory criterion of intelligence, and on the effective possibility to have a machine passing the test is abundant and controversial. At various stages of the history of artificial intelligence, various tasks have been posed to machines, and at times that the machine passed the test, questions were subsequently raised about the extent to which this "success" implied true intelligence.

Two familiar cases were Deep Blue, which was the first machine to win a chess game against a human champion,[53] and more recently Watson, which was the first machine to win Jeopardy!, a question and answer game, against human competitors. But the difference between Deep Blue and Watson in their way of processing information may help explain how our understanding of ways of producing intelligent behaviour has changed. Deep Blue was a mechanical brain equipped with enough logical knowledge and calculating power to be ahead of its opponent (however good) in foreseeing possible states of the game, thereby anticipating the adversary's possible (best) moves. Watson is a language processor, which is able to calculate the relevance of words in the hint (the form of questions in the game Jeopardy!) and has a vast database to search in order to produce a question (the form of answers, in that game) that it considers the most appropriate in the given context. There is of course a big difference between a manifestation of logical intelligence, combined with brute computational force, and one that is based on evaluating the relevance and interpretation of the many texts needed to be searched through. Each illustrates a typical way in which humans might be considered to act intelligently, though none of these ways can be taken to be an exclusive characterisation of intelligence. It seems appropriate in these cases to refer to the LoA (level of abstraction; see Chapter 2) at which we are considering intelligence. Similar examples show that machines have been able to face a certain task, and solve it appropriately, showing some form of intelligence, at a certain level. Besides the variety of levels at which human intelligence can be shown to act, such as logical, intuitive, semantic and analogical, there are more profound problems at stake here. Is a machine acting intelligently necessarily intentional? Is it necessary to ascribe semantic content to the source code (program) of a machine that seems to manifest intelligent behaviour?

It is curious that robots built to engage in experiments, such as the Turing Test, often perform very poorly, and that shows an insufficient pseudo-semantic behaviour when interacting with human users.[54] This is in striking contrast to the focused performances of machines, such as Deep Blue and Watson. The thesis that intelligence manifests itself at different levels, each presenting a certain way to process the relevant information relative to tasks and methods, gives us a far larger scope of possibilities to explore what intelligence is. However, it also explains why machine intelligence is still far away when measured in terms of interaction with human intelligence. If the intelligence of a mechanical system can be tested operationally in terms of LoAs on information, then it is not surprising that machines perform well at executing algorithms rapidly and selecting data, possibly, by probabilistic methods (i.e. tasks performed at the instructional level). On the other hand, what we identify as intelligent action in humans is, for example, accomplished by a method of goal-directed treatment

[53] Today, every appropriately equipped machine would always be able to beat many non-professional players at chess.
[54] Have fun exploring conversations with one such bot, the well-known ELIZA at http://www.masswerk.at/elizabot/.

of semantic information, and its counterpart methods of avoiding misinformation i.e. tasks performed at the semantic LoA.

The connection between intelligence and information does not only (at least partially) explain such great difference of results, but also supports the failure of the so-called equivalence fallacy (Copeland, 2000). According to the Equivalence Thesis, a universal Turing machine always has the potential to be any other system of the same class, and hence it can become a generally intelligent system in that class (Newell, 1980). The truth of the premise was established by Turing where he showed that a universal Turing machine can simulate any other specific Turing machine (and in general any other digital system up to the same computational power). However, the conclusion of the argument is doubtful in the general sense when intelligence is coupled with the ability to process information at different LoAs. Provided that a mechanical computation will deal primarily with instructional information, its ability to process semantic information in a significant way becomes a more problematic task, for which the appropriate ability to manipulate meanings (rather than just meaningless symbols) is needed.

13.7 Conclusion

We have seen that the relation between computation and information is undeniable. Various computational models, both digital and analogue, can be understood properly in terms of them being information manipulators. Some form of computation (at the very least, encoding and decoding procedures) is needed for conveying information. As well, in the absence of information, at least when construed as data or instructional information, there is very little one can do with computation. It is common to find the view in cognitive science discourse that computation, analogue or digital, is equivalent to information processing. Connectionists, who believe that neural networks are analogue computers (see Chapter 10), typically assert that these networks compute by virtue of information processing. Some computationalists, who believe that digital computers are adequate models of cognition, typically assert that these computers process information in a manner analogous to human cognisers.

Whether digital computation entails the processing of either Shannon information or factual information (and vice versa) is a deep and interesting question. It seems, at least prima facie, that digital computing systems are not up to the task of evaluating or ensuring truth, and if so, digital computation does not entail the processing of factual information. This is reminiscent of the symbol-grounding problem (see Chapter 6): digital computing systems only manipulate symbols and cannot reach out to the meaning of these symbols, let alone whether symbolic expressions (say, propositions) represent true state of affairs. As for Shannon information, it only makes sense in the context of a set of potential messages to be communicated and a probability distribution over this set. So, it seems problematic to argue that digital computation entails the processing of Shannon information. For digital computation can be either deterministic or not deterministic. Clearly, we cannot do justice here to assessing this problem.[55] But what this problem teaches us is that before we assess the claim that digital (or analogue) computation entails the processing of information, we need to understand what is meant by "information". Making such a claim, on the basis of a broad construal of information, leads to the paninformationalist view, according to which everything physical is information-theoretic in origin (see Chapter 3).

For our purposes, the main objective has been to illustrate the intimate relation between information and computation. Analogue computers operate on continuous physical signals and their counterparts operate on

[55] For more discussion see, for example, (Fresco, forthcoming; Piccinini & Scarantino, 2011).

discrete physical signals. This relation is fundamental for algorithmic information theory (see Chapter 14) and likewise important in the interactive process of questions and answers that is at the heart of the Turing Test, and the intelligent machinery which has since been developed in an attempt to pass this test. At this stage, we can only hope that our reader has in mind more questions than answers about computation.

13.8 Exercises

1. Starting from the axiom establishing that zero is a natural number ($0 \in N$) and the successor function (if $\in N; n+1 \in N$), one generates the set of natural numbers. The set of all arithmetical functions (multiplication, subtraction, power...) over integers is obtained by simple definitions over the basic set of computable functions (known as primitive recursive functions). Assume you have all such functions and can define any arithmetical operation. Now, using the latter, define new functions:

   - $f(n)$ such that takes natural numbers as objects and it returns the whole set of even numbers;
   - $f(n)$ such that takes natural numbers as objects and it returns the whole set of odd numbers;
   - $f(n)$ such that takes natural numbers as objects and it returns the whole set of prime numbers.
   - $f(n)$ such that takes the set of numbers 1-7 corresponding to weekdays as objects and it returns Monday-Wednesday-Friday as options chosen by Alice to take a train.

2. A Turing Machine can be described as a mechanical device that is able to compute all the recursive functions. Given the assumptions on the exercise above, a Turing Machine will be able to compute any of the arithmetical functions described above. Consider now such a Turing Machine as able to do the following: starting with an empty tape it can write 1s or 0s on it (W), it can scan 1s and 0s already printed (S), and it can move left (L) or right (R). Assume now your Turing Machine starts by the following instruction $S(1_1, 1_2)$, which means that it reads a symbol 1 on the first square of the tape and a symbol 1 on the second square of the tape. This is the way the Machine represents the numeral 2. The operations of the Machine are described by a list of instructions of the form $I(n_x, m_y, \dots)$, where $I = \{W, S, L, R\}$, $(n, m) = \{1,0\}$ for the possible symbols on the tape and $(x, y) = \{1,2, \dots\}$ for the numbered squares of the tape, so that such list gives the full behaviour of the Machine when executing a certain operation. Now define lists of instructions such that:
   - Starting from $S(1_1, 1_2, 1_3, 0_4, 1_5, 1_6)$ it finishes with $S(1_1, 1_2, 1_3, 1_4, 1_5)$
   - Which arithmetical operation has the machine implemented?

3. You can now generalize to a machine whose instructions implement functions of the form $F(n_x, m_y)$, where $F = \{+, -, \times, x^a\}$, $(n, m) = 2^{\{1,0\}}$ i.e. each argument is a set of 1s (any of them possibly empty) and the two arguments are separated by at least one 0; and $(x, y) = 2^N$ i.e. each argument takes n number of squares, where n=the number of 1s in each argument. The machine will end in a state $R(n)$, where $n = (nFM)$ i.e. reading the output of applying the relevant function to the two arguments. Can you define such machines for the functions $f(n)$ of one argument defined in the previous chapter?

4. Which examples can you name of systems (mechanical or human) displaying behaviour that can be defined as "intelligent" according to a description of actions at different levels of abstraction?

13.9 Further reading

Boolos, Burgess, and Jeffrey (2002), Hodges (1989), Petzold (2008), Primiero (2007), Primiero (forthcoming), White (2008).

Part VI: Special topics

Chapter 14

14. ALGORITHMIC INFORMATION THEORY

Quantifying simplicity and randomness (Beta Chapter)

14.1 Introduction

'Plurality is not to be posited without necessity.' William of Occam

If you have ever wondered if it makes any sense to ask whether a square can be said to be more or less complex than a squiggle, then algorithmic information theory is the theory that will allow you to answer, or at least to formulate, that question. Traditionally, ordinary information theory quantifies information by asking how many bits are needed to encode and communicate a *message* in a series of yes/no questions (bits). For example, it takes one bit to encode a single yes or no. (See introduction to Shannon in Chapter 1.) Algorithmic Information Theory (AIT) is of a very different character than ordinary information theory. AIT's motivation is the question *what does it mean for a sequence of 0s and 1s to be random?* It provides an answer by focusing on the lengths of computer programs that describe or reproduce a sequence. For example, a finite sequence of bits (also called a string) with a repeating pattern can be more succinctly described. In plain English one may say that the string 0101010101 can be described as "zero and one five times", while a more random-looking string would require a longer description, perhaps requiring spelling it out bit by bit.

For ordinary information theory, epitomized by Shannon's information entropy (see Chapter 1), the information content of a sequence is the number of bits needed to quantify each of the different sub-sequences (of any length) contained in a *message*. For algorithmic information theory, however, it is the minimum number of bits needed to store a program that produces the sequence that constitutes the information content of the sequence. Shannon's concept of entropy mirrors a concept in a field of physics called statistical mechanics, where Boltzman's entropy is the central founding concept. Just like Shannon's, Boltzman's entropy counts the number of different microstates constituting a physical macrostate, such as the number of different particle arrangements in a room full of gas. If the gas is all concentrated in a small space, the entropy of the room is low, because the number of different particle arrangements is small compared to the larger volume of the room. However, as soon as the constraint is released and the particles of the gas start filling the room, the entropy of the room increases because the particles can be arranged in more places. However, the gas will eventually fill the entire room and one can statistically describe the state of each particle inside. This is why the field was called statistical mechanics.

Algorithmic information theory, on the other hand, draws heavily on the theory of computation as initiated by Alan Turing. Developed independently by Ray Solomonoff, Andrei Kolmogorov and Gregory Chaitin, with crucial contributions from Leonid Levin, Per Martin-Löf, Claus Schnorr, Peter Gács and Charles Bennett among others, AIT attempts to quantify concepts such as randomness, simplicity and complexity that otherwise would remain informal or undefined. In this way, AIT can help to formulate and tackle questions in the Philosophy of Information (PI) that traditional tools from information theory are poorly equipped to deal with. Unlike ordinary information theory, AIT can deal with some questions related to semantics and epistemology. More precisely, it can shed light on the limits of formal knowledge that can be mechanically or effectively accessed (or not), about the known and unknown, and also the knowable and the unknowable.

For example, the string 011010011001011010011001101001… may look random when viewed through the lens of Shannon's entropy measure, but the sequence is simply generated by starting with 0 and then successively appending the "Boolean complement" (1 where there is a 0, or 0 where there is a 1) of the sequence obtained. Another look at the string may reveal this simple process 0,1,10,1001… where starting from 0 the next bit is 1, and the next two bits are thus the complement of 0 and 1, that is 10, and so on. Despite the infinite length of this sequence (known as the "Thue-Morse sequence"), there is a very short program of fixed length that can generate every bit, analogous to the short description we just gave in English. Such a program, however, would never be taken into consideration by Shannon's entropy, which would assign it a large information entropy value because the number of different substrings keeps growing, in spite of the fact that the generating program remains the same small size.

Although not everything is a bit string, bit strings are commonly chosen to illustrate AIT because they are simple, and because any message written in any alphabet can be rewritten in binary (although it is not always sensible to do so!). Returning to the example of the square and the squiggle, does it make sense to ask which shape is simpler? It does, because one can think of a minimal computer program producing the square, and producing the squiggle, and then ask which program is shorter. AIT will suggest that the simpler shape is the one produced by the shorter program. We know that a square can be described easily in a programming language. In fact, many programming languages already have a special purpose function to print squares and rectangles. But even writing a new function should not be difficult, as a square can be determined by a single number, specifying the height and the width. On the other hand, it is hard to simplify the description of a squiggle: probably the only way to describe it accurately to someone else would be to show them the squiggle itself. In some strong sense, therefore, one can expect a squiggle to require a longer computer program to be reconstructed, and therefore can be said to be more complex, or more random, than a square. Now you can see that although not everything is a bit string, in the end both the square and the squiggle can be generated by programs that are themselves bits inside a computer. So in general we will stay in the world of bits to describe AIT even though the claims of AIT are applicable to almost any object, no matter if it is an image, a graph or the blueprint of a car.

Shannon's important contribution of discovering that information could be quantified, using the concept of the bit, with no consideration of its semantic content (e.g. whether it is a car or a bike) is approached differently by AIT. Unlike the common belief that quantitative theories of information cannot shed any light on questions related to meaning, some of the concepts advanced by AIT are strongly related to epistemology. One of these concepts is the concept of comprehension, which has on several occasions been paired with various algorithmic information concepts, notably data compression. The theory of learning and optimal prediction can also be said to be settled in some formal sense by algorithmic

probability. Algorithmic probability is usually regarded as a formalization of a common paradigm in science, also known as "Occam's razor". One popular version says that '*among competing hypotheses, the hypothesis with the fewest assumptions should be selected',* although it was originally stated by William of Occam as 'Pluralitas non est ponenda sine necessitate' ('Plurality is not to be posited without necessity'). Chaitin's Omega number (see below) has also been called the "infinite wisdom number", which contains all knowledge but whose content is unattainable.

It seems that to compare descriptions of objects to determine which has the shortest description, one would need to specify some sort of "universal language" for this notion of "the shortest description" to make sense. Otherwise one could always come up with a language in which any arbitrary message had a short encoding, no matter how random it first appears. For example, Alice and Bob could agree to compress all the content of the Encyclopaedia Britannica in a single bit, so when Alice presents Bob with that bit, Bob could unfold it into the whole Encyclopaedia Britannica. AIT would achieve little if the complexity of something could be determined so arbitrarily. It is thus necessary to find a general language to "describe" any object in a fair fashion. This fair language, suggested by AIT, is the language of computer programs. The reason is that computer programs not only "rename" things, but they fulfil the extra requirement of giving a description that can actually reconstruct the original object, no matter the recipient. Now one can see why the Encyclopaedia Britannica cannot be paired with a small "descriptor", not to mention a single bit. Even Bob would need a computer program that already describes the Encyclopaedia Britannica to print all the volumes in their original form from the bit he is given by Alice.

14.2 Plain Kolmogorov complexity

Kolmogorov (1965) formally defined what is today known as the *plain* version of the Kolmogorov complexity (**K)** of a binary string, as the length of the shortest computer program (in bits) running on a universal Turing machine (that we will call **U**) that produces the string and halts (see Chapters 1 and 13). The Kolmogorov complexity, **K**, of a string such as 0101010101, is defined as the length in bits of the program printing out the string. In this case it can be produced by a short program implementing a printing loop of 01 five times. Any file in a computer is ultimately represented by bits, disregarding whether it is an image or a music file, or whether one measures its length in bytes or gigabytes (1 byte is 8 bits). Hence the range of application of **K** to any data is very large and is not limited to binary strings in any way. A string is said to be algorithmically random if the shortest program producing the string is the same length as the original string; on the contrary, the shorter the program that can produce the string, the less Kolmogorov random the string is.

As simple as it may sound, **K** is extremely powerful. It can be proven that **K** is able to "see" any possible regularity occurring in a dataset **s**. Finding regularities can be important for many reasons. One is for compressing data and saving precious resources by marking the regularity and producing instructions that reproduce it but occupy less memory space. One way to see **K(s)** is thus as providing the greatest possible compression rate for a perfect *lossless compression algorithm* compressing an object **s**. "Lossless" here means that the compressed version contains all the information of the original object, so that when decompressed the original object is fully recovered. (On information loss, see Chapter 3.) If an object has any type of pattern or regularity, **K** will make use of it to minimize the length of the computer program needed to describe it. A fair compression algorithm can be defined as one that transforms a string into two parts: one is the compressed version of the object, and the other the instructions to decompress the string, together accounting for the final length of the compressed object. In other words, one is adding

the decompression algorithm to the compressed string so that the compressed string comes with its own decompression instructions. Kolmogorov complexity establishes the ultimate limits of compression. It tells us why files cannot be compressed beyond certain limits and why they cannot be compressed to zero length (unless the file is empty).

K can be very useful for all sorts of applications related to semantics and meaning, in spite of the fact that it treats information in a formal and quantifiable fashion. Imagine Alice gives Bob the sequence 0101010101 and asks him if he believes it is the output of tossing a coin (where 0 is heads and 1 is tails). Would Bob believe the sequence is fair or random? He would certainly not, even though classical probability theory would say that 0101010101 has exactly the same chance of occurring than any other sequence of the same length. Classical probability theory does not help explain Bob's suspicion of this apparently non-random sequence. What about Shannon's information entropy? (See Chapter 1.) Shannon's information entropy would say that 0101010101 has a lower entropy than, say, 0110010100, but Shannon entropy cannot distinguish between 0101010101 and 000000011111, which may look equally suspicious to Bob. In fact, if Alice shows Bob a sequence such as 314159… and asks Bob whether the string looks random, Bob might recognize the 314159… digits as the first few digits of the mathematical constant π. Then this sequence does not look at all random to Bob, because it is generated by a simple process. Shannon's entropy, however, would assign maximum entropy to the digits of π, because π is believed to be a normal number, which is a number that contains all possible sub-sequences of digits. For example, all sub-sequences of 3 digits such as 123, 234 or 098 would occur exactly the same number of times in π, and the same for any other sequence of any length. This means that the Shannon entropy of the digits of π tend to maximal entropy for longer sequences of digits of π. However, π would have a very low Kolmogorov complexity because π is the result of an algorithmic process that can be described in a very concise way as the result of dividing the circumference of any circle by its diameter. For AIT, π would not be random, just as Bob thought when he recognized the sequence. Even when an approximation of **K(π)** would be difficult to estimate in practice, perhaps with a compression algorithm, **K** is equipped precisely to see this kind of structure and discriminate it from randomness, in a way that classical probability theory or other measures such as Shannon information entropy will not. For a general compression algorithm to see π, it would need to be "clever" enough, and, like Bob, it would need to have some knowledge about maths.

Now one can see how AIT helps understand meaning in a formal way in this context. Meaning for Bob with respect to the digits of π means that a bell rang in Bob's head when looking at π because he recognized that the digits have some meaning (ultimately, this is so even if Bob ignores it or fails to remember it, because π is a precise geometrical relation among all circles). The fact that Kolmogorov complexity can recognize π as a special object compared to, say, 0101010101 or 000000011111, seems to capture the idea of meaning and even of subjective meaning. Both the compression algorithm and Bob, if he had no knowledge of π, would fail to recognize π, but still they may eventually be able to see that π is the result of a simple, and highly compressible, calculation.

14.3 The founding theorem

If the language of computer programs is to be the language used to measure the complexity of an object, one would wish to have some stability under changes of programming language for **K(s)** to have general application. A theorem guarantees the stability of **K** using different programming languages (or universal Turing machines). One way to think about the "invariance theorem", as it is called, is to think of a program that translates between two programming languages or Turing machines $\mathbf{U_1}$ and $\mathbf{U_2}$ (U for Universal Turing machines; see Chapters 1 and 13). Because any Turing machine, particularly a universal one, can be implemented on any other universal Turing machine, the minimum length of a program on $\mathbf{U_1}$, plus the length of the translation program for $\mathbf{U_2}$ machine, will give the length of the program on $\mathbf{U_2}$. Think of a bilingual dictionary. The length of the dictionary is fixed and it may depend on $\mathbf{U_1}$ and $\mathbf{U_2}$ (the languages, in this case), but once we decide which language and which dictionary to use, the length of the dictionary remains the same for any word **s** in the dictionary, no matter how large or short **s** is. This invariance theorem is the foundation stone that established the field of AIT and is due to Solomonoff (1964), Kolmogorov (1965) and Chaitin (1966). The theorem establishes that there is a constant **c** that limits the difference of 2 measures as measured on 2 different Turing machines (or programming languages) $\mathbf{U_1}$ and $\mathbf{U_2}$, that is for all **s**, $\mid \mathbf{K_{U1}(s) - K_{U2}(s)} \mid \leq \mathbf{c}$.

The invariance theorem implies that the difference **c** between one measure of **K** under one computer language (or Turing machine) and another computer language is constant because it is the length in bits of the translator (or compiler) between $\mathbf{U_1}$ and $\mathbf{U_2}$ that matters and remains constant. The theorem says that **c** can be relatively large and have an impact if **s** is too small (small enough for this to happen), making the measurement of **K** potentially unstable for a while. However, as soon as **s** is longer than **c,** then the instability of **K** measured over different programming languages starts to become less and less important, eventually converging to the same **K** values. For example, imagine that under a programming language **U** the bit string **s**=0101010101 could be encoded by a program of length 10 bits, that is $\mathbf{K_U(s)=10}$, but for a programming language T it may take a program of length 15 bits to reproduce the same string, that is $\mathbf{K_T(s)=15}$. If the difference between $\mathbf{K_U(s)}$ and $\mathbf{K_T(s)}$ for all **s** is a constant **c=5** bits that can be determined by writing a translator between **U** and **T**, then this means that for **s** the difference between $\mathbf{K_U(s)}$ and $\mathbf{K_T(s)}$ is proportionally larger than for a string **s'=**01…01 (100 times larger than **s**). Then, the resulting difference between $\mathbf{K_U(s)}$ and $\mathbf{K_T(s)}$ for **s'** will be relatively smaller than for **s**, because it will remain at about **c=5** bits.

The theorem introduces a question, despite the positive meaning of the theorem. While the theorem guarantees the convergence of values when the length of **s** grows, the theorem says nothing about the rate of convergence. Indeed, **c** could be incredibly large, rendering the theory useless for short strings. Applications and recent reports suggest, however, that **c** is usually small enough in most *natural* cases to be neglected. The remaining philosophical question then is why in practice **c** remains small, and what does it mean for a computer language or a computer to behave in a *natural* way.

One common way to approximate **K** is with the use of lossless compression algorithms such as those used in common computer formats such as *gzip* and *rar*. The usefulness of lossless compression algorithms as a method for approximating **K** derives from the fact that compression is a sufficient test of non-randomness. The lossless compressed length of an object **s** (e.g. a string) is therefore an upper bound on **K(s)**, which means that while one cannot ever tell when a string is not compressible, if one succeeds

in somehow shortening a string, one can tell that its algorithmic complexity cannot be larger than the compressed length.

Now we can see why this discussion is important. It is important because the Kolmogorov complexity of an object **s, K(s),** is related to how compressible **s** will be, and it is not only a matter of whether **s** is random or not, but whether it is relatively random or not, and if the relativisation is about the length of the constant **c** then one cannot tell much about the string's randomness. The more regularities in **s,** the greater the compression rate. Because the compressed version of **s** together with the decompression instructions can be seen as a program compressing **s**, then **K(s)** is small if **s** is highly compressible. However, if the object is random, it also means it has no regularities and therefore it will be incompressible and will have a large **K(s)** value, in which case **s** will be said to be random (the less compressible, the more random). More generally, it is said that a string is **c**-incompressible or **c**-Kolmogorov random (or just **c**-random) if $\mathbf{K(s) \geq |s| - c}$.

14.4 Most salient properties of K

One observation from the definition of **K** is that most objects are random and are thus strings of maximal Kolmogorov complexity; we can see this in the following counting argument. First notice that the length of programs for **K** is always given in bits, and the strings themselves are given in bits. That both units are in bits is convenient in order to compare objects of the same type, but one can use any other base or convention. Nevertheless, most objects can be proven to be maximally random according to their generating computer programs. There are exactly $\mathbf{2^n}$ bit strings of length **n**, and there are $\mathbf{2^0 + 2^1 + 2^2 + \ldots + 2^{n-c} = 2^n - c}$ bit strings of length **n - c** bits. It follows then that there are considerably fewer short programs than long programs. Thus, the number of strings of length **n** that can be paired with a program of **c** bits shorter vanishes exponentially. One can't pair off all **n**-length binary strings with binary programs of much shorter length, because there simply aren't enough short programs to encode all strings in shorter strings, even under optimal circumstances. In fact, by the same argument, it is clear that among all the strings of certain length, there is always one string that cannot be compressed at all, not even by a single bit. Take as an example all strings of up to length 3. There are 14 such strings: 0, 1, 00, 01, 10, 11, 000, 001, 010, 101, 100, 011, 110 and 111, for which there are the same number of programs of the same length, and only six of shorter length (0, 1, 00, 01, 10 and 11). Thus most of the strings will be paired to programs of the same size and will have maximal Kolmogorov complexity. In other words, there are not enough bit programs of shorter length to pair all bit strings.

Another salient property of **K** is also commonly seen as its greatest burden. That is its uncomputable nature. A function is uncomputable if there is no Turing machine that is guaranteed to produce an output for its inputs, or in other words, if the machine computing the function doesn't halt for a number of inputs. For Kolmogorov complexity that means that the function $\mathbf{s \rightarrow K(s)}$ has no effective procedure (or Turing machine). That is, there is no general function that, given a specific string, can generate the shortest program that produces that string. The fact that Kolmogorov complexity cannot be computed stems from the fact that we cannot compute the output of every program, hence the halting problem (see Chapter 13).

This uncomputability of the function $\mathbf{s \rightarrow K(s)}$ is, however, also the source of its greatest strength. Contrary to the common belief that the greatest burden of **K** is its uncomputability, it is its uncomputability that provides **K** with its great power. AIT proves that no computable measure will be up to the task in finding all possible regularities among all possible infinite sequences. This is because there is

an uncountable number of possible regularities, while there is only a countable number of possible Turing machines (or computer programs), so there are not enough of them to spot every possible regularity. However, it is more precise to refer to the uncomputability of the function **s** → **K(s)** as semi-computability, because one can actually approximate **K(s)** from above i.e. one can calculate the upper bounds of **K**. One traditional way to calculate upper bounds on **K** is with the use of compression algorithms. A trivial upper bound on **K** for any string **s** is simply the program **print(s)**. If a string **s** does not allow any other shorter program than **print(s)** then **s** is said to be incompressible or algorithmically random.

A proof of the uncomputability of **K** is sometimes explained using Berry's paradox. The Berry paradox is a self-referential paradox arising from an expression of the type "consider 'the smallest positive integer not definable in fewer than twelve words'". Note, however, that the previous sentence has eleven words so that number could be defined in less than twelve words! Bertrand Russell (1906) was the first to discuss the paradox in the literature and he attributed it to G.G. Berry, a librarian at the Bodleian library in Oxford, who had suggested a similar but more limited paradox from the expression "the first undefinable ordinal". The paradox with these phrases is that while any set of integer numbers that has some property also has an integer smaller than any other that has the property in the set (by what is traditionally called the "well ordering principle"), one should be able to find an integer not definable in less than twelve words. However, the sentence itself is of less than twelve words and is "defining" the integer; hence such an integer cannot be defined. The resolution of the paradox is by making precise the word "definable". In the case of the measure **K**, "definable" was replaced by Chaitin (1995) with a computer program, and by doing so the sentence with **K** does not lead to a paradox but to a proof that **K** cannot actually be calculable. Or, put another way, if it were possible to compute the Kolmogorov complexity of any string, then it would also be possible to generate paradoxes of the Berry type systematically with **K** (that is descriptions shorter than they are supposed to be, rendering the notion of **K** meaningless). Both Chaitin's approach and a version by George Boolos (1989) using the Berry paradox lead to a proof of Godel's Incompleteness Theorem in a different and simpler way. The proof can be informally summarized by what is called "Chaitin's heuristic principle", that is, that axiom systems cannot prove the complexity of a formula if the formula has greater algorithmic complexity than the combined algorithmic complexity of the axioms in which the formula is intended to be proven.

14.5 A convenient variation of K

One interesting question is how often a string can be produced by a computer program whose instructions are picked uniformly randomly. One way to formulate the question is to ask for the probability of a string being produced by a universal Turing machine running a random program. However, the sum of the probabilities of all the computer programs that can produce a particular string is greater than 1 because there is an infinite number of computer programs that can produce the same string. The problem is that it is very easy to generate an infinite number of programs as an extension of a program that already produces a string. Take the program that prints **s** and then prints an extra 1 at the end, only to delete it before halting. The new program prints **s**, but it is only a spurious variation of a more compact program. For a measure to be called a probability, however, the sum of the probabilities has to be 1.

To circumvent this problem Leonid Levin (1974) and Gregory Chaitin (1977) devised a way to consider only significant programs. These are programs that are not initial subprograms of any other valid program. These types of sets are called "prefix-free domains". A classic example is the set of all telephone

numbers. The only way to reach a person by calling their telephone number is if no substring of its telephone number is a substring of any other telephone number. If Alice's number is 0123456789 and Bob happens to have 0123 as his telephone number, then Alice would never be reachable because every time anyone tries 0123456789 the telephone company would connect to 0123, to Bob. Different ways to avoid this are possible; for example, if the telephone company enforced a longer time delay or a special character as an indication of a telephone number termination (for example, some online banking systems ask customers to use the # sign to indicate termination). A more practical way to do this for the telephone number system is simply to require all telephone numbers in the world to be of the same length, including country and area codes. Alternatively, if a shorter version of the phone number exists locally, the shortcut would never be part of the initial segment of any other telephone number.

Prefix codes are guaranteed to exist for a countable set and the so-called Kraft (or Kraft-Chaitin) inequality guarantees that taking the sum of all the probabilities of the series will converge to 1, which is the necessary condition for a probability measure. To differentiate this variation of **K** we will denote it by **C**. Algorithmic (also known as Kolmogorov-Chaitin) complexity can now be rewritten as $\mathbf{C}_U = |\mathbf{p(s)}|$. Where **U** is now required to be a universal Turing machine that only accepts self-delimited programs **p**, and **U** is called a prefix-free universal Turing machine.

Most properties of **K** are inherited by the prefix-free variation **C**, such as its uncomputability and invariance, and hence we will invariably talk about **K** or **C** unless an explicit distinction is made. In fact, the difference between the two can be exactly quantified given that the number of extra bits that are needed to delimit an input string is small. By convention, when talking about **K**, we mean that the assertions apply both to **K** and **C**, but when talking about **C** it usually means that assertions only apply to **C**.

14.6 Optimal predictability and algorithmic probability

Algorithmic probability says that it is not the case that a single bit is the most complex random string, but actually the most structured possible one and, more importantly, that the complexity transition is smooth, more in accordance with intuition.

It may be that it makes sense that a single bit can be regarded as both the most simple and the most complex of strings from different perspectives, and the advantage of the algorithmic probability approach is that it provides not only a different notion of the complexity of a single bit (one that is in keeping with intuition), but also that it generates a different outcome to the compressibility approach, even when the two measures are intimately related and asymptomatically produce the same results in the long term (for longer strings). The two views reflect different aspects of what a single bit represents.

There is a measure **m** that describes the probability of a universal Turing machine producing a string **s** when running a computer program produced at random. **m** provides a distribution over the set of all strings that is known as the "Universal Distribution", and its properties have even been described as miraculous in the literature. The notion behind **m** is intuitive and powerful. If one wished to produce the digits of π randomly, one would have to try time after time until one managed to hit upon the first numbers corresponding to an initial segment of the decimal expansion of π. The probability of success is extremely small: 1/10 digits multiplied by the desired quantity of digits (for example, $(1/10)^{2400}$ for a segment of 2400 digits of π). But if instead of shooting out random numbers one were to shoot out computer programs to be run on a digital computer, the result would be very different. A program that

produces the digits of π would have a higher probability of being produced by a computer program. Concise and known formulas for π could be implemented as short computer programs that would generate any arbitrary number of digits of π. This measure can be written as follows:

$$\mathbf{m(s)_U = \sum_{U(p)=s} 1/2^{|p|}}$$

Where **|p|** is the length (in bits) of the programs that produce a string **s** running on a universal Turing machine **U**. In order to work **U** has to fulfil a minimal technical requirement, viz. that no valid computer program is the beginning of another valid computer program. It should be noted that the largest term in the sum of equation 1 is obtained when the denominator is the smallest, that is, when **|p|** is the smallest, namely the shortest length of program **p** in bits that produces **s**, but the length of the shortest program is nothing else but **C(s)**. Hence we have found the precise beautiful connection between frequency of production of a string and its algorithmic complexity.

The implications for the real world are broad and fascinating, if one speculates about the world as an unfolding algorithmic process. Suppose that all phenomena in nature can be carried out by a Turing machine as a computation. Then **m(s)** would be the algorithmic probability of an event **s** actually happening. In fact, **m(s)** can be used to explain the generation of structure out of randomness (random programs) at the smallest scale, generating order in the universe out of nothing.

14.7 Complexity and frequency

Algorithmic probability theory says that a string with low complexity (e.g. a repetition of "01" a hundred times) will be produced by a large number of random computer programs according to algorithmic probability. According to the Coding theorem, the string will also have a very short program among the programs producing it, short with respect to the string length when growing the string by repeating the same pattern. This in turn can be interpreted in the following way. If one is presented with a string that is assumed to be an ever-growing sequence from a source, asking what digits will come after the repetition of "01" a hundred times is determined, according to algorithmic probability, by the bits that preserve the (low) original Kolmogorov complexity of the string with the repetition of "01" a hundred times. In this case the answer is that algorithmic probability will say that the pattern "01" will be repeated again and again because it is the pattern that mostly preserves the original length of the generating computer program. So while the string "01" **n** times can grow very fast, the generating program will only grow very little, by about $\mathbf{\log_2 n}$.

Now suppose that 0 represents sunlight and 1 represents night; one can encode the days of a year as a sequence of 365 digits repeating the pattern 01, if one does not live near the earth's poles. Now one could formalize the question of whether the sequence of 01 will continue repeating after the 365th day. Classical probability alone would say that anything could happen (that a night can come after a night with no sunlight the next year, or that 10 sunlight days will come in a row followed by 3 nights). Instead, we see day after day a repeating pattern due to the movement of the Earth with respect to the Sun that will more likely repeat than stop. Ways to complement classical probability theory are traditionally of Bayesian nature, which gives some weight to previous observations. One can more naturally justify a repeating pattern and an expectation by way of algorithmic probability. In this case, the program behind a sequence of this type is governed by the laws of gravitation describing the rules of movement establishing the regular movement of the earth with respect to the Sun. Algorithmic probability is in this sense an optimal predictor that shows how theoretically reflecting on these measures and trying to estimate numerical

values can provide useful applications and understanding. Another application of algorithmic complexity can be subjective randomness testing, that is, how good humans are at perceiving and generating randomness. For example, most educated humans will favour sequences containing even and prime numbers because they may believe that even numbers look less random. Classical probability theory would say this is wrong, because odd, even or prime numbers are not more special than others. Algorithmic probability would say that producing one type of number is less random and may better quantify the failing. As it turns out, humans are poor random number generators in either case, but aspects could not be objectively quantified before.

We have seen how **m(s)** relates to **C(s)**, given that according to its definition, **m** obtains the greater summand from the shortest program that produces **s**. The Coding theorem further formalises the reverse relationship between **C(s)** and **m(s)**, establishing that **C(s)** is about **– log m(s)**. The theorem indicates that the algorithmic complexity of a string **s** is very close (up to the additive constant) to the negative value of the logarithm of the frequency of **s**. It tells us that if a string **s** is produced by many programs, then there is also a short program that produces **s**. Not by complete chance, this relation resembles Shannon's information entropy formula, but unlike Shannon's this other measure is related to a true measure of complexity in the sense that it is equipped to capture any regularity in **s** from inheriting the power of **K**, even though it means **K** is more difficult to estimate than Shannon's entropy. One question of great importance is what distribution of computer programs is assumed for calculating **m(s)**. While it is true that **K(s)**, unlike Shannon's entropy, is independent of probability distributions, this is no longer the case for **m(s)**. A nice property of **m(s)** and the reason it is often referred to as the Universal probability distribution is because it can be proven that **m(s)** dominates any other probability measure of computer programs. This is something similar to the invariance theorem, meaning that if another **m(s)** is defined in some other way, the new measure will end up behaving exactly as **m(s)** would do. The measure **m(s)**, however, assumes that the distribution of random computer programs is uniform, that is that every instruction in the computer program has equal chance of occurring. This assumption is a source of criticism as it makes a priori assumptions that can be thought of to be of the same type as traditional Bayesian approaches, invoking uninformative *prior* distributions (how a distribution looks or will look) that require caution.

In this way **m(s)** and the Coding theorem reconnect **K(s)** to classical probability theory by means of the question of the distribution of computer programs, from the assumption that programs of the same length are equally likely and it might be argued that in real-world situations this may be not the case. In defence of the approach, however, one can argue that a uniform distribution is the simplest non-informative distribution according to the principle of indifference; that is, if **n** possibilities are indistinguishable then each possibility should be assigned an equal probability of **1/n**.

14.8 The infinite wisdom number

We now move to the subfield of AIT that studies the algorithmic randomness of infinite objects where little, if any, real-world applications have been developed, but a rich foundational discussion arises. Algorithmic probability and the associated universal distribution function **s→m(s)** are closely related to another crucial concept in the theory of algorithmic information, that is, the Chaitin halting probability (Chaitin, 1975), also known as Chaitin's **Ω** (omega) number defined by $\Omega_U = \sum_{(p)\ halts} 1/2^{|p|}$. In this formula, every program **p** running on a (prefix-free) Turing machine **U** that halts contributes to the sum of the values $1/2^{|p|}$. The longer the program length $|p|$, the smaller the value of $1/2^{|p|}$, and therefore the smaller the contribution to **Ω**. Short programs, however, contribute largely to the most significant values of an **Ω** number. Just like **K** and **C**, **Ω** is also semi-computable, meaning one can estimate it (from below) by fixing a programming language framework and running random programs.

Knowing the first **n** bits of **Ω** would enable you to decide whether or not each program up to **n** bits in length ever halts, so knowing all digits would enable you to decide the halting problem of all possible computer programs. That is why **Ω** can be seen as encoding all possible answers to any computable question. Given that one can always formulate questions in terms of whether a Turing machine will halt (*yes* and *no* answers), one would have the answers to all mathematical questions. One could think of an **Ω** number as an oracle reminiscent of the answer given by the Deep Thought computer to the Ultimate Question of Life, the Universe, and Everything in the Douglas Adams's *The Hitchhiker's Guide to the Galaxy* science fiction series. But just as in this science fiction story where the computer gave the answer "42", answers given by an **Ω** number would be hard to understand and, in principle, impossible to follow. And again, just like in Adams's story, one would need to rely on another more powerful computer to verify the answer, which in turn may provide a more puzzling and impossible-to-follow answer.

Chaitin's **Ω** is in fact a family of numbers because its digit expansion depends on the chosen universal Turing machine **U** or programming language. For every chosen **U** there is a different **Ω** number. As a probability measure the summands in **Ω** should not add up to more than 1 and therefore **U** is also required to be a prefix-free Turing machine (or a self-delimited programming language), as was also the case for **m(s)**. Because **Ω** can never be 1, given that a number of Turing machines will never halt, **Ω** is more precisely called a semi-probability measure (as is **m(s)**).

Indeed, Solomonoff-Levin's semi-probability measure (or "semi-measure") $\mathbf{m_U(s)}$ provides an approximation to Ω_U, for the same **U**, together with the frequency value of the strings produced by the random programs that halt, and from which the string algorithmic probability (and the string Kolmogorov complexity **K(s)** from the application Coding theorem) can be estimated. Also just like **m(s)**, **Ω** is lower semi-computable, because one can numerically estimate lower bounds on both measures. Semi-computable measures like **Ω** and **m(s)** are also said to be computably enumerable (often shortened by c.e.).

When thinking of **Ω** in terms of a wisdom number containing infinite knowledge, including the answers to all questions that can be formulated as a computer program (e.g. all open mathematical problems and more), it is very interesting to find out that the digits of **Ω** are unattainable and incompressible, meaning that there are no shortcuts to reach that knowledge. No process can overrun **Ω** because it cannot be derived by any means simpler than the sequence of bits in **Ω** itself. That doesn't mean one cannot calculate a few digits of **Ω** for a number of cases. For example, if we knew that computer programs 0, 10

and 110 all halt (notice they are prefix-free), then we would know that the first digits of **Ω** are 0.111, which in turn, if we had started with 0.111 from this **Ω** number, we would know that the programs 0, 10 and 110 halt. In this sense, **Ω** encodes and maximally "compresses" the information of the halting state of all possible computer programs. Therefore, by knowing **Ω** one could solve the halting problem (see Chapter 13).

The problem is that the first **n** digits of **Ω** cannot be computed using a program significantly shorter than **N** bits long, which is also the basis of Chaitin's incompleteness theorem using AIT, based on the idea that a set of axioms cannot prove a theorem containing more information than the combination of axioms themselves. In other words, **Ω** is itself not computable, because a fixed length program would only be capable of estimating a few finite digits of the infinite number **Ω**. Although **Ω** has a simple mathematical definition, the definition does not enable us to determine more than a finite number of its digits and **Ω** numbers have been constructed for which no bit can be determined, even with the full power of set theory, which is the most powerful classical theory commonly used in mathematics. However, it is known that some **Ω** numbers are easier than others. Remember there can be an **Ω** for every (self-delimited) programming language or (prefix-free) universal Turing machine that changes the digits and order of the bits in **Ω**. This could be thought of as reformulating the question in another programming language. The interesting thing is that it is also known that there is an **Ω** number (or a family of them) that is maximally complicated in the sense that none of its digits can be computed, not even one. So even when one can reformulate questions in the form of computer programs in order to try to get some information from **Ω**, not only will most questions remain unanswered, but in some cases (for an unfortunate question) no answer can be extracted from some **Ω** numbers.

Algorithmic information theory can therefore make a surprising contribution to the philosophical discussion of the origins and limits of knowledge. According to Chaitin, AIT reveals that certain mathematical facts are true for no reason, a discovery that goes against Leibniz's principle of sufficient reason, but along with Aquinas' dichotomy of knowable and unknowable knowledge (that is knowledge or truths that can be known and have an explanation and knowledge or truths that are so for reasons that cannot be knowable or attained by any effective means).

Indeed, as it turns out, from the **Ω** number an infinite number of mathematical facts are irreducible in the sense that no effective theory (computation) can explain why they are true; that is, the only way to "prove" such facts is to assume them as axioms, that is assumptions that do not need any explanation at all and are assumed to be true by definition. But even assuming new axioms, there is always an infinite number of facts left that cannot be proven true or false. In this sense Chaitin claims that, unlike the common belief that mathematics is strange to randomness, mathematics is in a profound sense random as it contains facts that are true for no (computable) reason.

14.9 Convergence in definitions

One may ask whether all these measures can be taken seriously as fully characterising the intuitive notions of complexity and randomness once and for all. The most surprising result in AIT is what is known as the "convergence in definitions". This is a phenomenon similar to the convergence in definitions of the notion of algorithm in the 1930s when people such as Gödel, Church, Turing, Post, Kleene and others characterised the notion of an algorithm by different independent approaches that turned out all to be equivalent in computational power, giving the sense that the concept of algorithm had been definitely mathematically grasped by all these formulations. Something similar has happened with AIT. Kolmogorov, Chaitin, Levin, Schnorr and Martin-Löf have independently conceived different approaches to randomness (compression, predictability, typicality) that have turned out to be equivalent in various fundamental ways, particularly for the finite case (the convergence in definitions of the case of infinite sequences is debatable):

Incompressibility: we have explained how randomness can be characterised by incompressibility. According to Kolmogorov complexity, if an object is random then it is impossible to compress it. Incompressibility is a sufficient test for non-randomness; however the converse is not necessarily true as the power of lossless compression algorithms may be limited.

Typicality: we have also examined the notion of a regularity and how one can take advantage of regularities in order to compress an object. The basic idea is to spot the places where the regularity occurs and then code it with a shorter marker. One can devise statistical tests for regularities, for example, a test for whether a sequence is made of an even number of 1s or whether the digits of a sequence are the digits of a mathematical constant (such as π) in binary. Random sequences can then be characterised by failing or meeting all possible (computable) tests. In fact, Martin-Löf proves there is a single (but uncomputable) universal statistical test for every possible regularity in a sequence.

Unpredictability: another characterisation of a random sequence is by way of unpredictability. Schnorr shows that it is impossible to make money by guessing the next digits in a random sequence when using a computable betting strategy. One can intuitively see that if there are no predictable patterns and no regularities can be spotted in a random sequence, one can come up with a strategy to predict any digits. If this is actually the case, then the sequence is said to be random.

The convergence in definitions means that each definition assigns exactly the same *randomness* as each of the other definitions. In other words, the extension of each definition is the same; they contain the same objects, hence strongly suggesting that each definition has proven itself (at least for the case of finite sequences). One can write this beautiful result in a compact manner as follows:

incompressibility ↔ unpredictability ↔ typicality

That is, something that is incompressible is unpredictable and is typical in a statistical sense. A series of *universality results* (both in the sense of *general* and in the sense of Turing universal, the latter concept being a version of the former (Kirchherr, Li, & Vitányi, 1997)) leads to the conclusion that the definition of random complexity is mathematically objective from the following results:

- Martin-Löf proves that there is a universal (but uncomputable) statistical test that tests for all computably enumerable statistical tests. His definition of randomness is therefore general enough to encompass all effective tests for randomness.
- Solomonoff and Levin prove that algorithmic probability is a universal and optimal learning strategy with no prior knowledge.
- Schnorr shows that a predictability approach based in Martingales leads to another characterisation of randomness, which in turn is equivalent to Martin-Löf randomness.
- Chaitin proves that an algorithmically uncompressible sequence is also Martin-Löf random.
- The confluence of all these definitions.

When this happens in mathematics it is believed that a concept has been objectively captured (in this case the concept of randomness). However, some have raised concerns for convergence arguments in AIT, because the number of equivalent characterisations, compared to characterisations of the concepts of algorithm or computation, are much fewer and perhaps less strong (see Porter (2012)).

Another concern about convergence arguments is raised by Georg Kreisel, suggesting a possible systematic error: that is, a collection of definitions that converge to the wrong extension. Regarding the Church-Turing thesis, Kreisel writes, 'Equivalence results do not play a special role, simply because one good reason is better than 20 bad ones, which may be all equivalent because of systematic error' (Kreisel, 1971). The source of the systematic error, Kreisel seems to suggest, may be, for example, induced by historical factors, where a group of researchers trying to tackle the same problem find similar solutions that turn out to be equivalent only because they had similar (or identical) starting points and backgrounds. Nothing seems to suggest this is the case for the notion of algorithm (believed to be captured by the concept of computability) or for the notion of randomness (believed to be captured by AIT). However, it is very interesting to analyse the foundations of every approach, compare them and discuss the implications.

14.10 Conclusion

Algorithmic Information Theory (AIT) can be classified into two large subfields. One is concerned with the study of the algorithmic complexity of finite strings and the other is concerned with the study of the algorithmic randomness of infinite sequences (or sets of real numbers). They are deeply connected and carry rich content worth discussing, but at the same time they display important disagreement and incompatibilities that are also fascinating. Is this because of some essential difference? Or is it that the disagreement is only transient?

The convergence in definitions in AIT has brought together characterisations of randomness that were not believed to be so deeply formally connected, such as unpredictability, compressibility and regularity. In this brief account of AIT we have explained it as a powerful theory that provides a set of universal measures of complexity for individual data objects, particularly strings and sequences. These are universal in the sense that they were not advanced for any specific purpose other than quantifying a notion related to complexity, and, more importantly, they are proven to characterise all possible features in data. A common property of AIT measures are their various forms of uncomputability, which we have explained should not be taken as a barrier to make use of them, either theoretically or experimentally. This is not only because some simple cases are actually computable, but more importantly because the measures are

not completely uncomputable, but are approachable with tools such as compression algorithms that are sufficient proof of the non-randomness of an object.

We have also substantiated the claim that the quantitative treatment of information and information content by AIT (in the form of randomness) should not be taken as disqualifying AIT from providing important insights into deep philosophical questions of information, including the semantics of information. We have seen that its computability is also AIT's greatest strength, which no computable measure can match, and that the algorithmic probability approach is deeply related to long-standing questions of inductive inference, epistemology and even psychology.

A question that the reader may raise is whether randomness equals complexity. The intuitive concept of complexity usually means order and structure, something which does not entail the concept of Kolmogorov complexity, which means randomness. This is a fair point, and there is another important measure of complexity that the reader should be aware of that builds upon Kolmogorov complexity to grasp the intuitive concept of structural complexity, which we plan to cover in future versions. For the moment we should have a better view of what mathematics can do to quantify the notions of simplicity and degrees of randomness.

14.11 Exercises

1. Can you intuitively sketch ways to connect the concept of meaning to unpredictability, incompressibility and lack of regularity?

2. Suppose you have 2 objects, one that you were able to compress highly and one that you couldn't. What can you tell about them? Does it mean for certain that one is not algorithmically random but the other is? How could you prove one or another case?

3. In what way would you say a living organism is less or more algorithmically complex than, say, a stone? Discuss.

4. Does AIT imply that if most objects are algorithmically random, because one cannot pair all objects to shorter programs, then most physical phenomena will also lack a concise formula?

14.12 Further reading

Li and Vitányi (2008), Zenil (2011), Zenil (2013).

.

Chapter 15

15. PERSONAL IDENTITY

Who am I? What am I? What makes me, me?

15.1 Introduction

> *'The broader thesis I shall defend is that ICTs [Information and Communication Technologies] are, among other things, egopoietic technologies or technologies of self construction, significantly affecting who we are, who we think we are, who we might become, and who we think we might become.' (Floridi, 2011d, p. 550.)*

There are many questions of personal identity, and they are not all the same. Asking "who am I?" might be seeking an explanation of what makes you an individual, different from all other persons. Asking "what am I?" may require an answer that defines what is meant by "person", giving conditions for anything to be a person. These two questions are different, and the third is different again: asking "what makes me, me, the same person all my life?" This is the traditional question of personal identity, the *persistence* or *diachronic identity* question.

In this chapter, we will look at several questions of personal identity, and consider whether some of them are more interesting than others. We will start from the traditional account of the problem in order to illustrate the identity problem. Then, we will look at four views of personal identity, which are still discussed in the current literature: Locke's psychological criteria and Olson's biological criteria, Schechtman's narrative account, and Floridi's 3Cs model. Ultimately, we will see how Floridi's 3Cs model melds with Schechtman's narrative account to give a rich framework for approaching multiple questions of personal identity from an informational perspective.

15.2 The traditional personal identity question

The traditional personal identity question, the persistence or diachronic (through time) identity question, arises because things change over time. What makes something the same thing, even though it changes? This is an ancient question, often introduced using the example of Theseus' ship (recorded by Plutarch). Theseus' ship is commissioned, built, and sails off on her maiden voyage. Theseus sails his ship for many years, and naturally things break, and are repaired. Every winter, the ship is hauled out of the water for a more thorough overhaul, and even weakened planks on the hull are replaced. Eventually, every part of the original ship has been replaced. Is it the same ship? What makes it the same ship? Suppose that while this has been going on, a Theseus enthusiast has been carefully collecting all the discarded damaged parts.

He has assembled them back into the original ship, to put in a seafaring museum. Is the ship in the museum Theseus' ship?

Persons also change through time. You have few surface properties in common with the baby that you once were. You're a lot bigger, and your eyes and hair might have changed colour. Deeper properties also change. Childhood skills, such as spinning a hula-hoop or doing yo-yo tricks, may disappear; while you gain new ones, like changing the nappy of a squirming baby, or fixing a boiler. Knowledge, too, can be acquired, and lost. What makes a person the same person through all these changes? That is the persistence question of personal identity, and much of the literature on personal identity seeks to answer it.

We will look at three views of personal identity, initially intended as answers to the persistence question. The first, and traditionally most highly-favoured answer, is that persons persist on the basis of some kind of psychological continuity over time. Locke is the most famous proponent of such a view, holding that a person is "a thinking intelligent being, that has reason and reflection, and can consider itself as itself, the same thinking thing, in different times and places" (Locke, 1894).

While many agree that psychological features are the most important characteristics of persons, and so the basis for their continuity through change, it is tricky to say exactly which psychological features matter. One possible choice is memory, as memories do seem to be important to our sense of self. The problem is that we don't remember many things that happen to us! Do you remember what you were doing at 11.23am last Tuesday morning? Do you remember your favourite coffee shop? You can remember many details about it, but do you remember every time you went there? Yet you are the same person who went there, every time. We forget most of the mundane things that happen to us, yet we are still the persons who do those mundane things.

Other psychological features also change: you may not love or like the same people and things all your life, and you both gain and lose skills and other kinds of knowledge. So the basis for personal identity would have to be a cluster of psychological features causally related through time. Perhaps if you begin to learn the piano, become good at playing the piano, enjoy playing the piano, then gradually decline in skill so you cannot play very well anymore and so don't enjoy playing any more, that is one of the psychological features relevant to your personal identity, in the normal kind of causal relationship over time. If you have children, suddenly a new and fierce emotion and focus comes into your life, which persists in their childhood, then matures and changes as they become adult and their relationship to you changes; that is the normal kind of causal relationship over time. If you remember the important happenings in your life, but forget doing the washing, and other frequent, mundane events, then that is the normal kind of causal relationship over time. This contrasts with, say, being brainwashed into believing you are a duck, which is not a normal causal relationship. But while it is easy to give examples of the "normal kind" of causal relationship over time, it is difficult to say what this is, and it might be different for different skills, attachments, memories, and so on.

A second well-known problem with this account is that it seems at least possible that more than one person could have precisely this kind of causal psychological continuity with you. It is possible for a person to survive with half of their brain removed. But suppose half of your brain is removed and successfully transplanted into another body. Now we might have two distinct people who are psychologically continuous with you as you are now! They are both you. But if there was one thing we

believed when we set out to give an account of the persistence conditions for persons, it was that one person can only ever be one person. This is the fission problem (Parfit, 1971).

In view of these problems, the second well-known view on personal identity is animalism. On this view, your persistence conditions are the persistence conditions of the animal you are (Olson, 1997). You are a whole functioning organism – your DNA, cells, organs, immune system, central nervous system and so on. What makes you the same person over time is the spatio-temporal continuity of your body. We can track that over time, and it can only ever be in one place at a time. For the fission case above, you are the original body, whatever happens to the half of your brain that has been removed.

This agrees with what we actually think of as persons, and how we actually track them. We decide whether someone is the same person by tracking bodily continuity. Think of how we use fingerprints, and DNA. We regard these as more secure than the testimony of someone who claims to know the person – testimony that may be mistaken, or paid for. But very few people think this is a good view of personal identity, because they think that surely the psyche is as relevant as, or more relevant than, the body.

This brings us to the third view, the narrative theory, which can be seen initially as a development, albeit a significant development, of psychological continuity theories. We will look at Schechtman's Narrative Self-Constitution View (NSCV) (Schechtman, 2010). Schechtman holds that we constitute ourselves as persons by understanding our own lives as narratives, constructing the story of our lives. Persons all think about their own lives, and construct ideas of themselves as persons, and their lives as coherent over time. We think about ourselves as governed by norms – moral "oughts" like "I ought not to kill", or value oughts like "I ought to phone my Mum", or aesthetic oughts like "wearing purple and orange together is just not good" – and we build these into our narratives. Indeed, Schechtman thinks that thinking of ourselves as governed by such norms requires an autobiographical narrative.

Note that persons don't actually have to *tell* the story of their lives, even inside their own heads, to count as having a narrative. The stories are largely implicit, although of course bits of them get told at different points: when meeting new people, catching up with friends – even in job interviews and on CVs! Note that a person's narrative might not be entirely coherent. People often seem to tell a slightly different version of their story to their mother than to their friends, or to their husband than to their colleagues.

On the narrative theory, it looks like the right kind of causal continuity for psychological characteristics to form part of your personal identity is incorporation into your narrative. And this is partly a matter of your own choice. So it can be different for different persons. Success at work might mean more to some, and be a more important part of their personal identity; family links to another; friends to another; playing the piano to another, and so on. And this kind of variation seems absolutely right: one person can be a merchant banker; another is a pianist who just happens to work as a merchant banker.

This concludes the brief summary of the traditional debate. Recall that the views presented here are supposed to be answers to the persistence or diachronic identity question. They are attempts to seek something which can be the basis of your persistence, even while you change over time. The issues raised in the traditional debate are still important nowadays, as they can take new forms. For example, there are platforms like DeadSocial or LivesOn that allow you to send tweets and Facebook messages after you kick the bucket, in an attempt to extend your persistence after death.

15.3 Other personal identity questions

Notice that the persistence or diachronic identity question is an absolute question, looking for an absolute answer. You can remind yourself of problems with asking absolute questions by reading Chapter 2. That is, the question seeks an answer to what makes you the same person over time, seeking the same answer irrespective of who is asking the question, or why. But who is asking the question, and why, is important.

For example, consider a building that was originally built as a hospital, but is now used as a school. We can ask the question, "Is it the same building now?" and this is the persistence or diachronic identity question. But its answer depends on the purpose for asking the question. If Alice is looking for a hospital in an emergency, then the building is of no use to her – it is not the same building for her purposes. If Bob is trying to pick up his niece from school, and knows the way to the old hospital, then the old hospital is of use to Bob as the location of the school he is looking for. For that purpose, it is the same building.

The same is true of Theseus' ship. If the person asking whether it is the same ship is Theseus, and what he wants to know is which ship he can safely sail off in without it sinking, or other ship-owners accusing him of theft, then the repaired ship is Theseus' ship. Someone interested in history, with no intention of going sailing, would probably be very happy with the reconstructed version of the original parts of the ship to be found in the museum. For their purposes, that is Theseus' ship.

This is the lesson of applying the method of levels of abstraction (see Chapter 2). We always interact with the world for a particular purpose. In light of this purpose, we pay attention to certain features of what we are interacting with – such as the location of a building or the seaworthiness of a ship – and ignore other features – such as whether the building has 500 windows or fewer, or whether the ship has red sails or blue. For PI, there are no answers to questions that are absolute in the sense that they demand an answer independent of the level of abstraction at which the question is asked.

For PI, the same is true of the persistence question for personal identity. The answer depends on who is asking the question, and why. Different features of you matter for different purposes. Naturally, you are interested in whether you are the same person! But your family, friends, employer or employees, the law, and even Bess your cat, also have an interest. They all need you to continue to behave consistently towards them. This insight can be applied to the question of whether your original body, with half your brain, or the new body, which now has the other half of your brain, is you. From the point of view of the two bodies, if psychological continuity is what they value, then they are both you, and there will be a terrible mess. Two bodies now think they have the same lover and family, the same job, and live in the same house. From the point of view of your loved ones, who value your behaviour, both bodies may well display a distressing familiarity of understanding, knowledge, skills, and even manner. But the original body with the well-loved face will almost certainly triumph. From the point of view of the law, both bodies will know many things that only you should know, like bank account details, and email passwords. But the fingerprints and DNA of the original body are likely to trump such factors. If you have been law-abiding, the old body will get all your assets. If you have been a well-known international criminal, and the new body gets to the numbered Swiss bank accounts first, the new body might do very well.

So just as for Theseus' ship and the hospital-turned-school, there is no absolute answer to the question of who is you, independently of the purpose of asking the question, which sets the features relevant to answering the question.

This doesn't mean that there is no interesting question of personal identity. The diachronic identity question can be debated, once an LoA is agreed. For example, diachroneity can be explained as the sum of synchronic identity snapshots on a timeline – see below. Further, if we feel there's some kind of philosophical puzzle, there may well be a real philosophical puzzle, even if our first attempt at making the question more precise doesn't lead us to the interesting answers we hoped for. This is an indication that we need to work our way through to a better, more precise question.

Floridi suggests a further question for PI: "What keeps the self together as a whole and coherent unity: continuously existing and coherently behaving at any given time?" This question is also ancient, as Plato's question in the *Republic* and the *Phaedrus*, where he presents the human psyche as having three parts that are often in conflict. There is the intellectual part which understands what is good for you, and can reason sensibly about health and nutrition and so on; the emotional but cooperative part; and the uncooperative animal passions part, which wants everything that is bad for you! Plato worries how one single person can be in conflict; can both want the chocolate and not want the chocolate – how can you disagree with yourself? Floridi calls this the "synchronic (at the same time) identity question" (Floridi, 2011d). It is the question of what makes a person a whole single person at one particular time. This distinguishes it from the diachronic identity question, which asks what makes a person the same person over time, and the two questions can be regarded as complementary.

15.4 Answering different questions: Floridi's 3Cs and NSCV again

Floridi calls his view of personal identity the "three membranes model" – being the corporeal, the cognitive and the consciousness membranes – referred to simply as the 3Cs view (Floridi, 2011d). Remember that the psychological and the biological views of personal identity are competing attempts to answer the persistence or diachronic identity question. PI is only interested in addressing the persistence question when that question is asked for a specified purpose. Alternatively, PI is interested in the synchronic identity question, and Floridi's view is a response to that.

For Floridi, what is interesting is how we separate ourselves off from the rest of the world at all. The ancient Greeks sought to explain change, seeing stability at a time, but struggling to understand change across time. Floridi is reacting to our growing understanding of science, that it is stability that most needs explanation, not change. In our modern understanding, change just happens, and tends towards creating disorder, or entropy. For example, things break, get dirty, and decay. Things holding themselves together and resisting decay is what takes effort. Specifically, it takes incoming energy. The human body takes a lot of work to keep itself alive, which is why you need to eat, drink, and breathe. Floridi thinks persons are special because they are separated off from the rest of the world through three stages, or membranes, to create something very sophisticated.

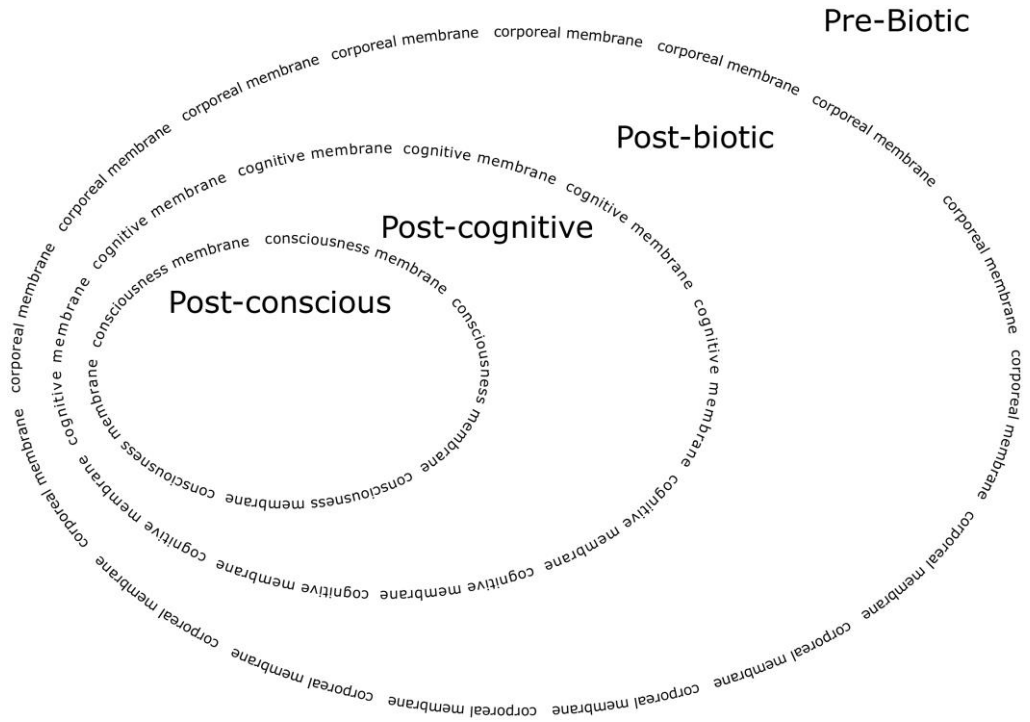

Figure 14: Floridi's 3Cs model

In figure 14, among pre-biotic structures there are no persons or even organisms. There are physical structures, that is, patterns of physical data. This is just the most general way of understanding everything that exists, using the very general concepts of information theory. For Floridi, information is the underlying notion of every level of abstraction used to perform the analysis. Then comes the first C, which concerns the evolution of organisms. In the post-biotic realm, some structures in the environment become contained in a corporeal membrane – such as in a body, contained perhaps by skin or a cell wall. The containment separates off the inside of the body or cell from the outside world. Floridi's example is a sunflower. The cognitive membrane is the second C. In the post-cognitive realm, data become encodable (in terms of Shannon's communication model), and organisms begin to be able to use data, such as sounds, visual patterns, gestures, smells, behaviours, and ultimately more sophisticated language. This requires a cognitive membrane, allowing the separation of data from the physical body for processing and communication, such as in memory. Floridi's example is a bird on the sunflower. The last C opens the door of the post-conscious or self-aware realm – i.e. where all information supported by consciousness can actually be constructed. Here, data become information and hence can be used for different purposes, and this includes assigning some data conventional meanings, such as making some sounds and words into a national anthem, which presumes a common, shared language and henceforth a public semantic. Floridi's example is a gardener watching the bird on the sunflower.

The language of the 3Cs view can be confusing for those who aren't already familiar with these ideas. But what's really fun is trying to work out what it implies. Filling out the core idea, using it to think with, is when we really get somewhere.

A nice way of doing that is connecting it with the narrative view. Schechtman says: 'According to NSCV, the limits of a person are determined by the limits of a narrative, and the integrity of a single person consists in the unity of a narrative.' (Schechtman, 2012, p. 336). First, the 3Cs view can be seen as giving the basis for being the kind of thing that can have a narrative, while also being the kind of thing that needs something to give it unity, including, ultimately, the continuous maintenance of that unity over time. We are not in control of our bodily or cognitive unity – these are gifts of evolution – but we can do something about our post-conscious unity. Then, instead of being an answer to the persistence question, the narrative view can be seen as giving an account of how the self-aware being goes about creating a person. Both views agree in respecting worldly constraints, seeing a person as continuously interacting with the world. So a narrative must respect the reality constraint, by conforming to basic facts, like people not living for 300 years, or being in two places at once; and the articulation constraint, by being able to be articulated locally, such as when someone asks how old you are, or about your family (Schechtman, 2012).

With the melding of the two views in mind, we can go on and look at two interesting issues for personal identity: personal identity online, and whether your personal identity is all about you.

15.5 Personal identity in an onlife world

For both the 3Cs model, and NSCV, persons in some sense constitute themselves, creating their own unity and their environment during this ongoing process matters. For example, in coming to be self-aware on the 3Cs model, you need to have a kind of successful separation between the information you have about the world, and the world itself – although you do want that information to be continuously influenced by the world! For NSCV, your environment needs to impact continuously on your self-narrative in the right way. If this is right, then the creation of a whole new world for persons to go and play in might well have interesting effects on personal identity. Note also that neither the 3Cs model, nor NSCV, say that there is anything essential to personal identity in being human, having a "normal" history or birth or childhood or body, or a soul. Floridi writes: 'If the self is made possible by the healthy development of all the three membranes, then any technology capable of affecting any of them is *ipso facto* a technology of the self.' (Floridi, 2011d, pp. 560-561). What this means is that since online technologies are capable of affecting your post-conscious membrane, and online relationships are capable of affecting your self-narrative, then the online world can affect your personal identity.

Floridi (2012a) argues that the rapid creation of the new world of the internet, the visible "infosphere", has drastically changed the whole picture: in the world based on Gutenberg, technologies were used to record and transmit data as auxiliary tools, while nowadays data are processed at such a level that human societies became dependent on information – they cannot work anymore without computing technology. How has this affected personal identity? Rodogno writes: 'Online contexts are novel and peculiar insofar as they afford prolonged disembodied and anonymous interaction with others.' (Rodogno, 2012, p. 309). Rodogno summarises the case for this affecting personal identity: first, it is easier to deceive online, so you may misrepresent yourself, and be deceived in turn; second, you cannot monitor others' reactions online in the way you can in person; third, entirely new possibilities are created online, such as the Facebook wall, which has no offline equivalent. You don't walk up to your friends in the bar and slap a message on a post-it-note on their foreheads. Schechtman considers this issue herself with respect to Second Life (SL), a sophisticated online game where players create avatars for themselves, and can spend many hours every day interacting with other avatars. For many players, the game is a fun way of exploring aspects of their personal identity they either can't explore offline, or are not ready to explore offline. For example, people can come out as gay in SL, long before they will tell offline friends. Or you can try out switching gender for a month! In these ways, SL can provide fascinating avenues for

exploring and choosing how to construct your own identity. Schechtman points out that SL and offline life interact with each other strongly. For example, SL avatars can design and make clothes, shoes, even skin, and build houses and so on in SL, and their offline users make offline money out of it. Couples can meet in SL and marry offline – and online. Offline couples have also divorced because of SL relationships, and successes and failures in SL affect offline sense of self, and vice versa. Schechtman thinks that sometimes the offline and SL narratives are both part of a broader, but single, person-narrative:

> sometimes the RL [real life] narrative of the user and the SL narrative of the avatar are, as it were, subplots in the more comprehensive narrative of the resident, a person who lives sometimes in RL and sometimes in SL. Both sets of adventures are part of the same life because, although distinguishable sub-narratives, they impact each other along the most fundamental dimensions of narrative interaction.
> (Schechtman, 2012, p. 341)

The Preface to the Onlife Manifesto summarises this interplay between online and offline with the new term *onlife* (Commission, 2013): now, the primacy of interactions rules over the primacy of entities i.e. the "Relational Self". If you don't interact with me, you do not exist. This leads to the concept of "freedom with elasticity", borrowed from economics: freedom exists in a complex environment where social constraints, technological artefacts, and nature defines our space of possibilities in acting as individuals in society. 'The broader thesis I shall defend is that ICTs [Information and Communication Technologies] are, among other things, egopoietic technologies or technologies of self construction, significantly affecting who we are, who we think we are, who we might become, and who we think we might become.' (Floridi, 2011d, p. 550.) Furthermore, the generational gap from digital immigrants to digital natives brings a different perspective on the relation between the physical world and the infosphere: for digital immigrants, they are ontologically separated – and so the personal identity is not defined by the ICTs actually used; on the contrary, digital natives will treat the physical world as immanent in the infosphere, so that any kind of separation will be perceived by the subject as dramatic, like a fish out of water (Floridi, 2014, p. 56).

And we see again why seeking an account of the unity of a person is such an interesting problem.

15.6 *My* personal identity: me, me, ME!

We will finish by considering who is asking the personal identity question about you, and why? Traditional treatments of personal identity make it all about you, while we see that in PI the emphasis is on you and others in some context: from solipsism to onlife. Durante (2011) sees the definition of personal identity as a trade-off between trust (where individuals should reveal something of themselves, based on cooperation) and privacy (where individuals pursue "freedom of", ultimately based on competition), but ICTs are changing the picture, so that digital natives implicitly seek "a balance between a traditional (based on settings and norms) and an informational idea of privacy (based on structural affordances)" (Durante, 2011, p. 613).

Wherever you are a digital immigrant or a digital native, you are not the only person with an interest in your personal identity. Your family, your friends, your employer and employees, your legal system and so on, all have rather strong interests in who you are – that you are the same person as yesterday, and in your unity or reliability, and so on. And you have a matching interest in all their identities.

We have seen that on the 3Cs model and NSCV, the environment of the person matters. And the most important environment of a person is other persons! And a lot of persons are using more and more ICTs to communicate and process information. Recall that although the self creates its own unity, and is continually

defining and redefining itself, the narrative must meet the worldly and articulation constraints, by respecting worldly facts, and being able to be articulated in parts, briefly and locally. Other persons and their relationships to you are important worldly facts, and constrain your ability to articulate your narrative. Floridi writes: 'We "identify" (provide identities) to each other, and this is a crucial (although not the only) variable in the complex game of the construction of personal identities' (Floridi, 2011d, p. 555.)

Because of this, other persons help create your narrative, and so your personal identity – for good or for ill. You might suffer a disappointment, and begin to see yourself as a failure, changing your narrative. But your friends and family might stubbornly refuse to react to you as a failure, continuing to celebrate your successes, and resisting your local "I'm a loser" articulations. Over time, they can alter your narrative, pushing it, and your life with it, back towards success. One of the interesting features of the combination of the 3Cs model and NSCV is that they show how important these kinds of experiences really can be to personal identity. On the other side of the coin, they help explain just why it is so exasperating when other people continue to articulate your narrative in a way that doesn't suit you – the teacher who remembers when you couldn't do ballet, and fell over during the school play to the amusement of absolutely everybody except you, or the ex who continues to believe, in spite of overwhelming evidence to the contrary, that you are still in love with him or her. Such people seize control of your own narrative, and attempt to re-write it. However temporarily, they are interfering with your identity. Of course it is annoying! But your friends are interfering too, when they refuse to see you as a failure.

For good or ill, though, we should expect such conflict. We write other people into our narratives, and get written into theirs. Embarrassing childhood experiences for you might be vital identity-constituting episodes for your parents, and they will not forget them. And we should expect the continuing growth of the infosphere to continue to affect our interactions with other persons, and so our identities. Floridi again: 'Online communities—understood as dynamic, interactive and distributed networks, in which the individual is never a stand-alone entity but always a participant—play a vital role in the creation of PIO [personal identity online].' (Floridi, 2011a, p. 478). And we have seen that a central part of that is the importance of our relationships to others, just as it has always been: 'The infosphere is not just a medium, but the new environment where groups and individuals continuously and increasingly define themselves.' (Floridi, 2011a, p. 478).

So with the melding of the 3Cs model and NSCV, we separate ourselves from the rest of the world, and we can reconnect with it in as many ways as we choose. This accounts for the variation in the answers to other questions of personal identity from person to person and from time to time. And other people are vital. They write themselves into our narratives, and us into their narratives. We can cooperate or resist, as we choose. Some people find this scary. Others find it liberating!

15.7 Exercises

1. What question of personal identity is of most interest to you, and why?

2. Do you think you create your own identity? How?

3. Do you think you could ever see a Second Life avatar, or some other online creation, as a seamless part of your real life?

4. Do you perceive yourself as a digital immigrant or native? How did your use of ICTs influence your answer?

5. How important are other people to your identity?

15.8 Suggestions for the exercises

1. For example, are you more interested in the question of what makes you the same person across time (diachronic identity), or what makes the aspects of your identity form a coherent whole at a particular time (synchronic identity)? What is your purpose in asking the question?

2. Think about whether you have a narrative. Does it constitute your identity? Or is it irrelevant to your identity? Do you have multiple narratives? Do they ever conflict?

3. Have a think about it. Perhaps have a go at it! Alternatively, there are various artists investigating the relationship between digital media and our identities. Have a look at Stelarc at http://stelarc.org/?catID=20247.

4. Think about your first use of ICTs to process information relevant for your personal identity. How old were you? Did your peers have similar experiences?

5. Think not only about the obvious candidates – family and close friends – but also daily acquaintances, colleagues, and even people you don't know at all, such as in the media.

15.9 Further reading

Floridi (2011d) and Floridi (2011a) give Floridi's arguments for the important constructive role of information technology for our personal identity. Floridi (2014) devotes Chapter 3 to how identity (at a personal and social level) is shaped in our world. Schechtman (2010) and Schechtman (2012) explore her narrative self-constitution view, and apply it to online contexts. Rodogno (2012) explores different questions of personal identity. Finally, the Onlife Manifesto (Commission, 2013) and the commentaries by scholars coming from different backgrounds is a reference for issues in personal identity and selfhood.

REFERENCES

Adams, F. (2005). Tracking theories of knowledge. *Veritas, 50*, 1-35.

Allo, P., & Mares, E. (2012). Informational semantics as a third alternative? *Erkenntnis, 77*(2), 167-185. doi: 10.1007/s10670-011-9356-1

Arquilla, J. (1999). Ethics and information warfare. In K. Khalilzad, J. White & A. Marsall (Eds.), *Strategic Appraisal: The Changing Role of Information in Warfare* (pp. 379-401). Santa Monica, USA: Rand Corporation.

Arquilla, J., & Borer, D. A. (2007). *Information strategy and warfare: A guide to theory and practice*. New York: Routledge.

Arquilla, J., & Ronfeldt, D. F. (1997). *In Athena's camp: Preparing for conflict in the information age*. Santa Monica, Calif.: RAND.

Baase, S. (2012). *A gift of fire: Social, legal, and ethical issues for computing technology* (4 ed.): Prentice Hall.

Bachelard, G. (2002). *The formation of the scientific mind: A contribution to a psychoanalysis of objective knowledge*. Manchester: Clinamen.

Baggini, J., & Fosl, P. S. (2007). *The ethics toolkit: A compendium of ethical concepts and methods* (1 ed.): Wiley-Blackwell.

Bar-Hillel, Y., & Carnap, R. (1953). Semantic information. *The British Journal for the Philosophy of Science, 4*(14), 147-157.

Barlow, H. B. (1961). Possible principles underlying the transformation of sensory messages. In W. Rosenblith (Ed.), *Sensory Communication* (Vol. 1, pp. 217-234): MIT Press.

Barwise, J. (1997). Information and impossibilities. *Notre Dame Journal of Formal Logic, 38*(4), 488-515.

Barwise, J., & Seligman, J. (1997). *Information flow: The logic of distributed systems*. Cambridge: Cambridge University Press.

Bateson, G. (1972). *Steps to an ecology of mind*: University of Chicago Press.

Baudrillard, J. (1976). *L'èchange symbolique et la mort (Symbolic exchange and death)*. [Paris]: Gallimard.

Baudrillard, J. (1981). *Simulacres et simulation (Simulacra and simulation)*. Paris: Editions Galilée.

Baudrillard, J. (1987). *L'autre par lui-même (The ecstasy of communication)*. Paris: Editions Galilée.

Baudrillard, J. (2001). *The vital illusion*. New York; Chichester: Columbia University Press.

Beall, J. C. (2008). *Spandrels of truth*. Oxford: Oxford University Press.

Beaney, M. (Ed.). (2012). *The Stanford encyclopedia of philosophy* (Summer 2012 ed.).

Beer, R. D. (2000). Dynamical approaches to cognitive science. *Trends in Cognitive Sciences, 4*(3), 91-99. doi: 10.1016/s1364-6613(99)01440-0

Ben-Jacob, E., Shapira, Y., & Tauber, A. I. (2005). Seeking the foundations of cognition in bacteria: From Schrödinger's negative entropy to latent information. *Physica A: Statistical Mechanics and its Applications, 359*, 495-524.

Black, T. (2006). Contextualism in epistemology. In J. Fieser & B. Dowden (Eds.), *Internet encyclopedia of philosophy* (July 15, 2006 ed.).
Blackburn, S. (2002). *Being good: A short introduction to ethics* (2nd Revised ed.): Oxford Paperbacks.
Boden, M. A. (2008). Information, computation, and cognitive science. In P. Adriaans & J. Van Benthem (Eds.), *Philosophy of Information* (pp. 741-761). Amsterdam: Elsevier.
Bogdan, R. J. (1988). Information and semantic cognition: An ontological account. *Mind and Language, 3*(2), 81-122. doi: 10.1111/j.1468-0017.1988.tb00136.x
Boolos, G. (1989). A new proof of the Gödel incompleteness theorem *Logic, logic, and logic* (pp. 383-388): Harvard University Press.
Boolos, G., Burgess, J. P., & Jeffrey, R. C. (2002). *Computability and logic* (4th ed. ed.). Cambridge: Cambridge University Press.
Bremer, M. E., & Cohnitz, D. (2004). *Information and information flow: An introduction*. Frankfurt: Ontos Verlag; Lancaster.
Brooks, A. (1991). Intelligence without representation. *Artificial Intelligence Journal, 47*.
Brown, B., & Priest, G. (2004). Chunk and permeate: A paraconsistent inference strategy. Part I: The infinitesimal calculus. *Journal of Philosophical Logic, 33*(4), 379-388.
Bush, V. (1931). The differential analyzer: A new machine for solving differential equations. *Journal of the Franklin Institute, 212*(4), 477-488.
Carnap, R., & Bar-Hillel, Y. (1952). An outline of a theory of semantic information. Cambridge, Massachusetts: MIT.
Castells, M. (2000). *The rise of the network society*. Oxford; Malden, Mass.: Blackwell Publishers.
Chaitin, G. J. (1966). On the length of programs for computing finite binary sequences. *Journal of the ACM, 13*(4), 547-569.
Chaitin, G. J. (1975). A theory of program size formally identical to information theory. *Journal of the ACM, 22*(3), 329-340.
Chaitin, G. J. (1977). Algorithmic information theory. *IBM Journal of Research and Development, 21*(4), 350-359.
Chaitin, G. J. (1995). The Berry Paradox. *Complexity, 1*, 26-30.
Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory*. Oxford: Oxford University Press.
Chemero, A. (2009). *Radical embodied cognitive science*: MIT Press.
Chomsky, N. (1957). *Syntactic structures*. Gravenage: Mouton & Co.
Chomsky, N. (1959). A review of B. F. Skinner's Verbal behavior. *Language, 35*(1), 26-58.
Church, A. (1936). An unsolvable problem of elementary number theory. *American Journal of Mathematics, 58*(2), 345-363.
Churchland, P. M. (1992). *A neurocomputational perspective: The nature of mind and the structure of science*: MIT Press.
Clark, A. (1998). *Being there: Putting brain, body, and world together again*: The MIT Press.
Clark, A. (2001). *Mindware: An introduction to the philosophy of cognitive science*. New York ; Oxford: Oxford University Press.
Clark, A., & Chalmers, D. J. (1998). The extended mind. *Analysis, 58*(1), 7-19.
Collins, J. M. (2006). Epistemic closure principles. In J. Fieser & B. Dowden (Eds.), *Internet encyclopedia of philosophy* (June 13, 2006 ed.).
Commission, E. (2013). The onlife manifesto: Being human in a hyperconnected era. Retrieved from http://ec.europa.eu/digital-agenda/futurium/sites/futurium/files/Manifesto.pdf
Copeland, B. J. (1993). *Artificial intelligence: A philosophical introduction*. Oxford ; Cambridge, Mass.: Blackwell.

Copeland, B. J. (1996). What is computation? *Synthese, 108*(3), 335-359. doi: 10.1007/bf00413693
Copeland, B. J. (2000). Narrow versus wide mechanism: Including a re-examination of Turing's views on the mind-machine Issue. *The Journal of Philosophy, 97*(1), 5-32.
Cordeschi, R. (2008). Cybernetics. In L. Floridi (Ed.), *The Blackwell Guide to the Philosophy of Information and Computing* (pp. 186-196). Oxford: Blackwell.
Corritore, C. L., Kracher, B., & Wiedenbeck, S. (2003). On-line trust: Concepts, evolving themes, a model. *International Journal of Human-Computer Studies, 58*(6), 737-758. doi: 10.1016/s1071-5819(03)00041-7
Council, A. (1992). ACM Code of ethics and professional conduct, 2012, from http://www.acm.org/about/code-of-ethics
D'Alfonso, S. (2011). On quantifying semantic information. *Information, 2*(1), 61-101.
De Tienne, A. (2006). *Peirce's logic of information*. Paper presented at the Seminario del Grupo de Estudios Peirceanos, Universidad de Navarra.
Deleuze, G. (1968). *Différence et répétition (Difference and repetition)*: Paris: Presses universitaires de France.
Deleuze, G. (1990). Post-scriptum sur les sociétés de contrôle (Postscript on the societies of control). *L'Autre journal, 1*.
Demir, H. (Ed.). (2012). *Luciano Floridi's philosophy of technology* (Vol. 8).
Denning, D. E. (1999). *Information warfare and security*. Reading: Addison-Wesley.
Doyle, A. (1985). Is knowledge information-produced belief? A defense of Dretske against some critics. *The Southern Journal of Philosophy, XXIII*, 33-46.
Dretske, F. (1970). Epistemic operators. *Journal of Philosophy, 67*, 1007-1023.
Dretske, F. (1981). *Knowledge and the flow of Information*. Cambridge MA: MIT Press.
Dretske, F. (1983). Precis of Knowledge and the flow of information. *Behavioral and Brain Sciences, 6*(1), 55-63.
Dretske, F. (2000). *Perception, knowledge, and belief: Selected essays*. Cambridge: Cambridge University Press.
Dretske, F. (2003). How do you know you are not a zombie? In B. Gertler (Ed.), *Privileged access and first-person authority*. Burlington: Ashgate.
Dretske, F. (2008). *Epistemology and information*.
Ducheyne, S. (2005). Joan Baptiste van Helmont and the question of experimental modernism. *Physis: Rivista Internazionale di Storia della Scienza,, 43*, 305-332.
Dummett, M. A. E. (1993). *Origins of analytical philosophy*. London: Duckworth.
Dunn, J. M. (2008). Information in computer science. In P. Adriaans & J. V. Benthem (Eds.), *Philosophy of Information* (pp. 581-608). Amsterdam; London: Elsevier.
Dupuy, J. P. (2000). *The mechanization of the mind: On the origins of cognitive science*. Princeton, N.J.: Princeton University Press.
Durante, M. (2011). The online construction of personal identity through trust and privacy. *Information 2011, 2*(4), 594-620. doi: 10.3390/info2040594
Dyson, G. (2012). *Turing's cathedral: The origins of the digital universe*. London: Allen Lane.
Epstein, R. G. (1996). *The case of the killer robot: Stories about the professional, ethical, and societal dimensions of computing*. New York ; Chichester: Wiley.
Ess, C., & Thorseth, M. (2011). *Trust and virtual worlds: Contemporary perspectives*. New York: Peter Lang.
Fallis, D. (2002). Introduction: Social epistemology and information science. *Social Epistemology, 16*(1), 1-4. doi: 10.1080/02691720210132752
Fieser, J., & Dowden, B. (2007). Epistemology. In J. Fieser & B. Dowden (Eds.), *Internet Encyclopedia of Philosophy*.
Floridi, L. (1999). *Philosophy and computing: An introduction*. London; New York: Routledge.

Floridi, L. (2003). On the intrinsic value of information objects and the infosphere. *Ethics and Information Technology, 4*(4), 287-304.
Floridi, L. (2004a). On the logical unsolvability of the Gettier problem. *Synthese, 142*(1), 61-79.
Floridi, L. (2004b). Open problems in the philosophy of information. *Metaphilosophy, 35*(4), 554-582.
Floridi, L. (2004c). Outline of a theory of strongly semantic information. *Minds and Machines, 14*(2), 197-221.
Floridi, L. (2005a). Consciousness, agents and the knowledge game. *Minds and Machines, 15*(3), 415-444.
Floridi, L. (2005b). Is semantic information meaningful data? *Philosophy and Phenomenological Research, 70*(2), 351-370.
Floridi, L. (2007a). In defence of the veridical nature of semantic information. *3*, 31-42.
Floridi, L. (2007b). A look into the future impact of ICT on our lives. *The Information Society, 23*(1), 59-64.
Floridi, L. (2008). Trends in the philosophy of information. In P. Adriaans & J. van Benthem (Eds.), *Handbook of Philosophy of Information* (pp. 113-132): North Holland.
Floridi, L. (2009). The information society and its philosophy: Introduction to the special issue on "The philosophy of information, its nature, and future developments". *The Information Society, 25*(3), 153-158. doi: 10.1080/01972240902848583
Floridi, L. (2010a). Information ethics. In L. Floridi (Ed.), *The Cambridge handbook of information and computer ethics*: Cambridge University Press.
Floridi, L. (2010b). Information, possible worlds, and the cooptation of scepticism. *Synthese, 175*(1), 63-88.
Floridi, L. (2010c). *The philosophy of information*. Oxford: Oxford University Press.
Floridi, L. (2010d). The philosophy of information as a conceptual framework. *Knowledge, Technology and Policy, 23*(1), 253-281. doi: 10.1007/s12130-010-9112-x
Floridi, L. (2010e). Semantic information and the correctness theory of truth. *Erkenntnis, 74*(2), 147-175.
Floridi, L. (2011a). The construction of personal identities online. *Minds and Machines, 21*(4), 477-479. doi: 10.1007/s11023-011-9254-y
Floridi, L. (2011b). A defence of constructionism: Philosophy as conceptual engineering. *Metaphilosophy, 42*(3), 282-304. doi: 10.1111/j.1467-9973.2011.01693.x
Floridi, L. (2011c). *Information: A very short introduction*. Oxford: Oxford University Press.
Floridi, L. (2011d). The informational nature of personal identity. *Minds and Machines, 21*(4), 549-566. doi: 10.1007/s11023-011-9259-6
Floridi, L. (2012a). Hyperhistory and the philosophy of information policies. *Philosophy and Technology, 25*(2), 129-131. doi: 10.1007/s13347-012-0077-4
Floridi, L. (2012b). Semantic information and the network theory of account. *Synthese, 184*(3), 431-454. doi: 10.1007/s11229-010-9821-4
Floridi, L. (2014). *The fourth revolution: The impact of information and communication technologies on our lives*: Oxford University Press.
Floridi, L. (forthcoming-a). Information closure and the sceptical objection. *Synthese*.
Floridi, L. (forthcoming-b). *Information ethics*: Oxford University Press.
Floridi, L. (forthcoming-c). Perception and testimony as data providers. *Logique et Analyse*.
Floridi, L., & Sanders, J. W. (1999). Entropy as evil in information ethics. *Etica and Politica, special issue on Computer Ethics, 1*(2).
Floridi, L., & Sanders, J. W. (2001). Artificial evil and the foundation of computer ethics. *Ethics and Information Technology, 3*(1), 55-66. doi: Doi 10.1023/a:1011440125207
Fodor, J. A. (1983). *The modularity of mind: An essay on faculty psychology*. Cambridge, Mass. ; London: MIT Press.

Fodor, J. A. (1997). Connectionism and the problem of systematicity (continued): Why Smolensky's solution still doesn't work. *Cognition, 62*(1), 109-119.
Foucault, M. (1966). *Les mots et les choses. Une archéologie des sciences humaines (The order of things. An archaeology of human sciences)*: [Paris].
Foucault, M. (1969). *L'archéologie du savoir (The archaeology of knowledge)*. [Paris]: Gallimard.
Fox, C. J. (1983). *Information and misinformation: An investigation of the notions of information, misinformation, informing, and misinforming*. Westport CT, U.S.A: Greenwood Press.
Frege, G. (1892). Über sinn und bedeutung. *Zeitschrift für Philosophie und philosophische Kritik, 100*, 25-50.
Fresco, N. (2010). A computational account of connectionist networks. *Recent Patents on Computer Science, 3*(1), 20-27.
Fresco, N. (forthcoming). Information processing as an account of concrete digital computation. *Philosophy and Technology*, 1-30. doi: 10.1007/s13347-011-0061-4
Freud, S. (1917). A difficulty in the path of psycho-analysis. *The Standard Edition of the Complete Psychological Works of Sigmund Freud, XVII*(1917-1919), 135-144.
Gambetta, D. (1998). Can we trust trust? In D. Gambetta (Ed.), *Trust: Making and breaking cooperative relations* (pp. 213-238): Basil Blackwell.
Gelder, T. v. (1998). The dynamical hypothesis in cognitive science. *Behavioral and Brain Sciences, 21*, 615-628.
Gelven, M. (1994). *War and existence: A philosophical inquiry*. University Park, Pa.: Pennsylvania State University Press.
Gettier, E. L. (1963). Is justified true belief knowledge? *Analysis, 23*(6), 121-123.
Gibson, J. J. (1986). *The ecological approach to visual perception*: Lawrence Erlbaum.
Giere, R. N. (2006). *Scientific perspectivism*: The University of Chicago Press.
Glanzberg, M. (2013). Truth. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy. Retrieved from http://plato.stanford.edu/archives/spr2013/entries/truth/.
Gleick, J. (2011). *The information: A history, a theory, a flood*. London: Fourth Estate.
Goldman, A. I. (1976). Discrimination and perceptual knowledge. *The Journal of Philosophy, 73*(20), 771-791.
Graham, G. (1999). *The internet: A philosophical inquiry*. London: Routledge.
Greco, G. M., Paronitti, G., Turilli, M., & Floridi, L. (2005). *How to do philosophy informationally?* Paper presented at the WM2005: Professional Knowledge Management, Kaiserslautern.
Grice, H. P. (1957). Meaning. *The Philosophical Review, 66*(3), 377-388.
Grice, H. P. (1975). Logic and conversation *The Logic of Grammar* (pp. 64-75). Encino, CA: Dickenson.
Grice, H. P. (1989). *Studies in the way of words*. Cambridge, Mass.: Harvard University Press.
Hacking, I. (1982). Biopower and the avalanche of printed numbers. *Humanities in Society, 5*, 279-295.
Harms, W. F. (1998). The use of information theory in epistemology. *Philosophy of Science, 65*(3), 472-501. doi: 10.1086/392657
Harnad, S. (1990). The symbol grounding problem. *Physica Scripta, D*(42), 335-346.
Harris, R. (2005). *The semantics of science*. London ; New York: Continuum.
Haugeland, J. (1997). *Mind design II: Philosophy, psychology, artificial intelligence* (Rev. and enl. ed.). Cambridge, Mass.; London: MIT Press.
Hemp, D. (2006). The KK (knowing that one knows) principle. In J. Fieser & B. Dowden (Eds.), *Internet encyclopedia of philosophy* (October 15, 2006 ed.).
Hetherington, S. (2005). Gettier problems. In J. Fieser & B. Dowden (Eds.), *The internet encyclopedia of philosophy*.

Hintikka, J. (1973). *Logic, language-games and information: Kantian themes in the philosophy of logic*. Oxford: Clarendon Press.
Hodges, A. (1989). *Alan Turing, enigma* (2. Aufl. ed.). Wien: Springer-Verlag.
Ihde, D. (1979). *Technics and praxis*. Dordrecht, Holland ; Boston: D. Reidel Pub. Co.
Johnson, D. G. (1985). *Computer ethics*. Englewood Cliffs: Prentice-Hall.
Kant, I. (2002). *Groundwork for the metaphysics of morals*. Oxford: Oxford University Press.
Kirchherr, W., Li, M., & Vitányi, P. (1997). The miraculous universal distribution. *Mathematical Intelligencer, 19*, 7-15.
Klein, J. (2008). Francis Bacon's scientia operativa, the tradition of The workshops, and the secrets of nature. In C. Zittel, R. Nanni, G. Engel & N. Karafyllis (Eds.), Philosophies of technology: Francis Bacon and his contemporaries: Brill E-Books. doi: DOI:101163/ej.9789004170506.i-582.1.
Klein, J. (2009). Francis Bacon. In E. N. Zalta (Ed.), The Stanford Encyclopaedia of Philosophy. Retrieved from http://plato.stanford.edu/archives/spr2009/entries/francis-bacon/.
Kolmogorov, A. N. (1965). Three approaches to the quantitative definition of information. *Problems of Information and Transmission, 1*(1), 1-7.
Kreisel, G. (1971). Some reasons for generalizing recursion theory. In R. Gandy & C. M. E. Yates (Eds.), *Logic colloquium 69* (pp. 139-198). Amsterdam.
Lakoff, G. (1987). *Women, fire, and dangerous things : what categories reveal about the mind*. Chicago: University of Chicago Press.
Leroi-Gourhan, A. (1964-65). *Le geste et la parole (Gesture and speech)*. Paris: Albin Michel.
Levin, L. A. (1974). Laws of information conservation (non-growth) and aspects of the foundation of probability theory. *Problems of Information Transmission, 10*(3), 206-210.
Lewis, D. (2004). Letters to Beall and Priest. In J. Beall & G. Priest (Eds.), *The law of non-contradiction: New philosophical essays* (pp. 176-177). Oxford: Oxford University Press.
Li, M., & Vitányi, P. (2008). *An introduction to Kolmogorov complexity and its applications* (3 ed.). New York: Springer.
Locke, J. (1894). *An essay concerning human understanding*. Oxford: Clarendon Press.
Luhmann, N. (1979). *Trust and power: Two works*. Chichester; New York: Wiley.
Luper, S. (2010). The epistemic closure principle. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2010 ed.): Stanford University.
MacKay, D. M. (1969). *Information, mechanism and meaning*. Cambridge, Mass.; London: MIT Press.
MacLennan, B. J. (2004). Natural computation and non-Turing models of computation. *Theoretical Computer Science, 317*(1-3), 115-145. doi: 10.1016/j.tcs.2003.12.008
Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*: W. H. Freeman and Company: New York.
Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3), 175-183. doi: 10.1007/s10676-004-3422-1
Maudlin, T. (1989). Computation and consciousness. *The Journal of Philosophy, 86*(8), 407-432.
McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biology, 5*(4), 115-133.
Mitcham, C. (1994). *Thinking through technology: The path between engineering and philosophy*. Chicago; London: University of Chicago Press.
Moor, J. H. (1985). What is computer ethics? *Metaphilosophy 16*(4), 266-275.
Moschovakis, J. (2010). Intuitionistic logic. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2010 ed.).

Mumford, L. (1934). *Technics and civilization*. U.S.A.: G. Routledge & Sons.
Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review, 83*(4), 435-450.
Nauta, D. (1970). *The meaning of information*: Mouton.
Newell, A. (1980). Physical symbol systems. *Cognitive Science, 4*, 135-183.
Newell, A., & Simon, H. A. (1972). *Human problem solving*: Prentice-Hall Englewood Cliffs, NJ.
Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM, 19*(3), 113-126.
Nissenbaum, H. (1998). Protecting privacy in an information age: The problem of privacy in public. *Law and Philosophy, 17*(5 - 6), 559-596.
Norton, J. (1987). The Logical inconsistency of the old quantum theory of black body radiation. *Philosophy of Science, 54*(3), 327-350.
Olsen, J.-K. B., Pedersen, S. A., & Hendricks, V. F. (2009). *A companion to the philosophy of technology*. Chichester: Wiley-Blackwell.
Olson, E. T. (1997). *The human animal: Personal identity without psychology*. New York; Oxford: Oxford University Press.
Osgood, C. E. (1952). The nature and measurement of meaning. *Psychological Bulletin, 49*(3).
Parfit, D. (1971). Personal identity. *Philosophical Review, 80*(1), 3-27. doi: 10.2307/2184309
Parker, D. B. (1968). Rules of ethics in information processing. *Communications of the ACM, 11*(3), 198-201. doi: 10.1145/362929.362987
Peirce, C. S. (1977). *Semiotic and significs: The correspondence between Charles S. Peirce and Victoria Lady Welby*. Bloomington: Indiana U.P.
Peirce, C. S. (1984). *Writings of Charles S. Peirce: A chronological edition* (Vol. 2 1867-1871). Bloomington: Indiana University Press.
Penrose, R. (1989). *The emperor's new mind: concerning computers, minds, and the laws of physics*. Oxford: Oxford University Press.
Petzold, C. (2008, 2012/07/13/21:59:43). The annotated Turing a guided tour through Alan Turing's historic paper on computability and the Turing machine, from http://proquestcombo.safaribooksonline.com/9780470229057
Piazza, T. (2010). Perceptual evidence and information. *Knowledge, Technology and Policy, 23*(1), 75-95.
Piccinini, G. (2008a). Computers. *Pacific Philosophical Quarterly, 89*(1), 32-73. doi: 10.1111/j.1468-0114.2008.00309.x
Piccinini, G. (2008b). Some neural networks compute, others don't. *Neural Networks, 21*(2-3), 311-321. doi: 10.1016/j.neunet.2007.12.010
Piccinini, G., & Scarantino, A. (2011). Information processing, computation, and cognition. *Journal of Biological Physics, 37*(1), 1-38. doi: 10.1007/s10867-010-9195-3
Popper, K. R. (1963). *Conjectures and refutations. The growth of scientific knowledge. (Essays and lectures.)*. London: Routledge & Kegan Paul.
Porter, C. (2012). *Mathematical and philosophical perspectives on algorithmic randomness*. PhD, University of Notre Dame. Retrieved from http://www3.nd.edu/~cholak/papers/porter12.pdf
Priest, G. (2006). *Doubt truth to be a liar*. Oxford: Clarendon.
Priest, G. (2008). *An introduction to non-classical logic* (2nd ed. ed.). Cambridge: Cambridge University Press.
Primiero, G. (2007). An epistemic constructive definition of information. *Logique et Analyse*(200), 391-416.
Primiero, G. (forthcoming). Offline and online data: On upgrading functional information to knowledge. *Philosophical Studies*, 1-22. doi: 10.1007/s11098-012-9860-4
Putnam, H. (1981). *Reason, truth and history*. Cambridge: Cambridge University Press.

Pylyshyn, Z. W. (1984). *Computation and cognition: Toward a foundation for cognitive science*. Cambridge, Mass: MIT Press.
Quine, W. V. (1960). *Word and object*. Cambridge, Mass.: MIT Press.
Read, S. (1995). *Thinking about logic: An introduction to the philosophy of logic*. Oxford: Oxford University Press.
Rey, G. (1997). *Contemporary philosophy of mind: A contentiously classical approach*. Cambridge, Mass.; Oxford: Blackwell.
Robin, R. S. (1967). *Annotated catalog of the papers of Charles S. Peirce* (Vol. 463). [S.l.]: University of Massachusetts Press.
Rodogno, R. (2012). Personal identity online. *Philosophy and Technology, 25*(3), 309-328. doi: 10.1007/s13347-011-0020-0
Russell, B. (1906). Les paradoxes de la logique. *Revue de métaphysique et de morale, 14*, 627-650.
Russell, B. (1948). *Human knowledge, its scope and limits*. London: George Allen & Unwin.
Russo, F. (2011). Correlational data, causal hypotheses, and validity. *Journal for General Philosophy of Science, 42*(1), 85-107. doi: 10.1007/s10838-011-9157-x
Russo, F. (2012). The homo poieticus and the bridge between physis and techne. In H. Demir (Ed.), *Luciano Floridi's philosophy of technology* (Vol. 8, pp. 65-81).
Ruyer, R. (1954). *La cybernétique et l'origine de l'information (Cybernetics and the origin of information)*. Paris: Flammarion.
Rysiew, P. (2011). Epistemic contextualism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2011 ed.).
Sackett, D. L., Rosenberg, W. M. C., Gray, J. A. M., Haynes, R. B., & Richardson, W. S. (2007). Evidence based medicine: What it is and what it isn't. *Clinical Orthopaedics and Related Research, 455*, 3-5.
Scarantino, A., & Piccinini, G. (2010). Information without truth. *Metaphilosophy, 41*(3), 313-330.
Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals and understanding: an inquiry into human knowledge structures*. Hillsdale, N.J.: Erlbaum ; New York ; London : Distributed by Wiley.
Schechtman, M. (2010). Personhood and the practical. *Theoretical Medicine and Bioethics, 31*(4), 271-283. doi: 10.1007/s11017-010-9149-6
Schechtman, M. (2012). The story of my (second) life: Virtual worlds and narrative identity. *Philosophy and Technology, 25*(3), 329-343. doi: 10.1007/s13347-012-0062-y
Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*(03), 417-424. doi: doi:10.1017/S0140525X00005756
Sequoiah-Grayson, S. (2007). The metaphilosophy of information. *Minds and Machines, 17*(3), 331-344. doi: 10.1007/s11023-007-9072-4
Serres, M. (1995). *Genesis*. Ann Arbor: University of Michigan Press.
Shannon, C. (1938). *A symbolic analysis of relay and switching circuits*. Masters, MIT.
Shannon, C. (1941). Mathematical theory of the differential analyzer. *Journal of Mathematics and Physics, 20*, 337-354.
Shannon, C. (1948). The mathematical theory of communication. *Bell Systems Technical Journal*.
Shannon, C. (1993). *Claude Elwood Shannon: Collected papers*. New York, N.Y.: Institute of Electrical and Electronics Engineers.
Shapiro, J. A. (2007). Bacteria are small but not stupid: Cognition, natural genetic engineering and socio-bacteriology. *Studies in the History and Philosophy of Biological and Biomedical Sciences, 38*, 807-819.
Shapiro, S. (2006). *Vagueness in context*. Oxford: Clarendon.
Shettleworth, S. J. (1998). *Cognition, evolution, and behavior*: Oxford University Press.
Simondon, G. (1958). *Du mode d'existence des objets techniques (On the mode of existence of technical objects)*: Paris.

Simondon, G. (1964). *L'Individu et sa genèse physico-biologique. L'Individuation à la lumière des notions de forme et d'information (Individuation in the light of the notions of form and information)*: Paris.
Skinner, B. F. (1938). The behavior of organisms: An experimental analysis.
Skinner, B. F. (1990). Can psychology be a science of mind? *American Psychologist, 45*(11).
Skinner, B. F. (1992). *Verbal behavior*. Acton, Mass.: Copley.
Skinner, B. F., & Ferster, C. B. (1997). *Schedules of reinforcement*: Copley Publishing Group.
Smolensky, P. (1986). Information processing in dynamical systems: Foundations of harmony theory. In D. Rumelhart (Ed.), *Parallel distributed processing: Explorations in the microstructure of cognition: Foundations* (Vol. 1): MIT Press.
Solomonoff, R. J. (1964). A formal theory of inductive inference: Parts 1 and 2. *Information and Control, 7*, 1-22; 224-254.
Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy, 24*(1), 62-77. doi: 10.1111/j.1468-5930.2007.00346.x
Sperling, G. (1963). A model for visual memory tasks. *Human Factors, 5*(1), 19-31.
Stalnaker, R. C. (1984). *Inquiry*. Cambridge, Mass.: MIT Press.
Steup, M. (2006). The analysis of knowledge. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2012 ed.).
Taddeo, M. (2009). Defining trust and e-trust. *International Journal of Technology and Human Interaction, 5*(2), 23-35. doi: 10.4018/jthi.2009040102
Taddeo, M. (2010). Modelling trust in artificial agents, A first step toward the analysis of e-trust. *Minds and Machines, 20*(2), 243-257. doi: 10.1007/s11023-010-9201-3
Taddeo, M. (2011). Information warfare: A philosophical perspective. *Philosophy and Technology, 25*(1), 105-120. doi: 10.1007/s13347-011-0040-9
Taddeo, M., & Floridi, L. (2005). Solving the symbol grounding problem: A critical review of fifteen years of research. *Journal of Experimental and Theoretical Artificial Intelligence, 17*(4), 419-445.
Tarski, A. (1944). The semantic conception of truth and the foundations of semantics. *Philosophy and Phenomenological Research, 4*, 341-376.
Thelen, E., & Smith, L. B. (1994). *A dynamic systems approach to the development of cognition and action*. Cambridge, Mass.; London: MIT Press.
Thomson, A. (1999). *Critical reasoning in ethics: A practical introduction*: Routledge.
Turilli, M., Vaccaro, A., & Taddeo, M. (2010). The case of online trust. *Knowledge, Technology and Policy, 23*(3-4), 333-345. doi: 10.1007/s12130-010-9117-5
Turing, A. M. (1936). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematics Society, 2nd series, 42*, 230-265.
Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433-460.
Vakarelov, O. (2010). Pre-cognitive semantic information. *Knowledge, Technology and Policy, 23*(2), 193-226.
Van Benthem, J., & Martinez, M. (2008). The stories of logic and information. In P. Adriaans & J. Van Benthem (Eds.), *Philosophy of Information*: North-Holland.
von Neumann, J. (1958). The computer & the brain. *Recherche, 67*.
Weckert, J. (2005). Trust in cyberspace. In R. J. Cavalier (Ed.), *The impact of the internet on our moral lives* (pp. 95-120). Albany: University of New York Press.
White, G. (2008). The philosophy of computer languages. In L. Floridi (Ed.), *The Blackwell guide to the philosophy of computing and information* (pp. 237-247). Oxford: Blackwell.
Wiener, N. (1948). *Cybernetics or control and Communication in the animal and the machine* (2nd ed. ed.). Cambridge, MA: MIT Press.
Wiener, N. (1988). The human use of human beings: Cybernetics and society: Da Capo Press.

Wilcox, S., & Jackson, R. (2002). Jumping spider tricksters: Deceit, predation, and cognition. *The Cognitive Animal*, 27-33.

Williams, P. L., & Beer, R. D. (2010). *Information dynamics of evolved agents*. Paper presented at the From Animals to Animats.

Wilson, R. A., & Foglia, L. (2011). Embodied cognition. In E. N. Zalta (Ed.), The Stanford encyclopedia of philosophy. Retrieved from http://plato.stanford.edu/archives/fall2011/entries/embodied-cognition/.

Wittgenstein, L. (1961). *Notebooks, 1914-1916* (G. E. M. Anscombe, Trans.): Oxford: Basil Blackwell.

Zenil, H. (Ed.). (2011). *Randomness through computation*: World Scientific Publishing Company.

Zenil, H. (Ed.). (2013). *A computable universe*: World Scientific Publishing Company.