This is a preprint version of the following article:
Brey, P. and Søraker, J. (2009). 'Philosophy of Computing and Information Technology' *Philosophy of Technology and Engineering Sciences. Vol. 14 of the Handbook for Philosophy of Science.* (ed. A. Meijers) (gen. ed. D. Gabbay, P. Thagard and J. Woods), Elsevier.

Philosophy of Computing and Information Technology

Abstract
Philosophy has been described as having taken a 'computational turn', referring to the ways in which computers and information technology throw new light upon traditional philosophical issues, provide new tools and concepts for philosophical reasoning, and pose theoretical and practical questions that cannot readily be approached within traditional philosophical frameworks. As such, computer technology is arguably the technology that has had the most profound impact on philosophy. Philosophers have studied computer technology and its philosophical implications extensively, and this chapter gives an overview of the field. We start with definitions and historical overviews of the field and its various subfields. We then consider studies of the fundamental nature and basic principles of computing and computational systems, before moving on to philosophy of computer science, which investigates the nature, scope and methods of computer science. Under this heading, we will also address such topics as data modeling, ontology in computer science, programming languages, software engineering as an engineering discipline, management of information systems, the use of computers for simulation, and human-computer interaction. Subsequently, we will address the issue in computing that has received the most attention from philosophers, artificial intelligence (AI). The purpose of this section is to give an overview of the philosophical issues raised by the notion of creating intelligent machines. We consider philosophical critiques of different approaches within AI and pay special attention to philosophical studies of applications of AI. We then turn to a section on philosophical issues pertaining to new media and the Internet, which includes the convergence between media and digital computers. The theoretical and ethical issues raised by this relatively recent phenomenon are diverse. We will focus on philosophical theories of the 'information society', epistemological and ontological issues in relation to Internet information and virtuality, the philosophical study of social life online and cyberpolitics, and issues raised by the disappearing borders between body and artifact in cyborgs and virtual selves. The final section in this chapter is devoted to the many ethical questions raised by computers and information technology, as studied in the field of computer ethics.

Table of Contents

1. Introduction

Philosophers have discovered computers and information technology (IT) as research topics, and a wealth of research is taking place on philosophical issues in relation to these technologies. The philosophical research agenda is broad and diverse. Issues that are studied include the nature of computational systems, the ontological status of virtual worlds, the limitations of artificial intelligence, philosophical aspects of data modeling, the political regulation of cyberspace, the epistemology of Internet information, ethical aspects of information privacy and security, and many, many more. There are specialized journals, conference series, and academic associations devoted to philosophical aspects of computing and IT as well as a number of anthologies and introductions to the field [Floridi, 1999, 2004; Moor and Bynum, 2002], and the number of publications is increasing every year.

Philosophers have not agreed, however, on a name for the field that would encompass all this research. There is, to be fair, not a single field, but a set of loosely related fields – such as the philosophy of artificial intelligence, computer ethics and the philosophy of computing – which are showing some signs of convergence and integration yet do not currently constitute one coherent field. Names considered for such a field tend to be too narrow, leaving out important areas in the philosophical study of computers and IT. The name "philosophy of computing" suggests a focus on computational processes and systems, and could be interpreted to exclude both the discipline of computer science and the implications of computers for society. "Philosophy of computer science" is too limiting because it suggests it is the study of an academic field, rather than the systems produced by that field and their uses and impacts in society "Philosophy of information technology", finally, may put too much emphasis on applications of computer science at the expense of computer science itself.

Without aiming to settle the issue for good, we here propose to speak of the area of *philosophy of computing and information technology*. We define philosophy of computing and IT as the study of philosophical issues in relation to computer and information systems, their study and design in the fields of computer science and information systems, and their use and application in society. We propose that this area can be divided up into five subfields, which we will survey in the following five sections. They are the philosophy of computing (section 2), the philosophy of computer science (section 3), the philosophy of artificial intelligence (AI) (section 4), the philosophy of new media and the Internet (section 5), and computer and information ethics (section 6). A reasonably good case can be made, on both conceptual and historical grounds, that these areas qualify as separate fields within the broad area of philosophy of computing and IT. Conceptually, these areas have distinct subject matters and involve distinct philosophical questions, as we will try to show in these sections. We also believe that these areas have largely separate histories, involving different, though overlapping, communities of scholars.

Historically, the *philosophy of AI* is the oldest area within philosophy of computing and IT.

Philosophy of AI is the philosophical study of machine intelligence and its relation to human intelligence. It is an area of philosophy that emerged in close interaction with development in the field of artificial intelligence. The philosophy of AI studies whether computational systems are capable of intelligent behavior and human-like mental states, whether human and computer intelligence rest on the same basic principles, and studies conceptual and methodological issues within various approaches in AI. The philosophy of AI started to take shape in the 1960s, and matured throughout the 1970s and 1980s.

The *philosophy of computing* is a second area that formed early on, and in which significant work was being done since at least the 1970s. As defined here, it is the philosophical study of the nature of computational systems and processes. The philosophy of computing studies fundamental concepts and assumptions in the theory of computing, including the notions of a computational system, computation, algorithm, computability, provability, computational complexity, data, information, and representation. As such, it is the philosophical cousin of theoretical computer science. This area, which is more loosely defined and contains much less research than the philosophy of AI, is the product of three historical developments. First, the philosophy of AI necessitated an understanding of the nature of computational systems, and some philosophers of AI consequently devoted part of their research to this issue. Second, philosophically minded computer scientists working in theoretical computer science occasionally started contributing to this area. A third factor that played a role was that philosophers working in philosophical logic and philosophy of mathematics started considering fundamental issues in computing that seemed to be an extension of the issues they were studying, such as issues in computability and provability of algorithms.

By the late 1980s, the landscape of philosophical research on computers and IT consisted almost entirely of studies on AI and theoretical issues in computing. But grounds were shifting. With the emergence of powerful personal computers and the proliferation of usable software, computers were becoming more than an object of study for philosophers, they were becoming devices for teaching and aids for philosophical research. In addition, philosophers were becoming increasingly concerned with the social impact of computers and with ethical issues. At several fronts, therefore, the interest of philosophers in issues relating to computers and computing was therefore increasing.

Playing into this development, some philosophers started advancing the claim that philosophy was gearing up for a “computational turn”, an expression first introduced by Burkholder [1992] and also advanced, amongst others, by Bynum and Moor [1998]; the argument was already advanced in the 1970s by Sloman [1978]. The *computational turn in philosophy* is a perceived or expected development within philosophy in which an orientation towards computing would transform the field in much the same way that an orientation towards language restructured the field in the so-called linguistic turn in twentieth-century Anglo-American philosophy. At the

heart of the argument for the computational turn was that computing did not just constitute an interesting subject matter for philosophy, but that it also provided new models and methods for approaching philosophical problems [Moor and Bynum, 2002].

The application of computational tools to philosophy, referenced by the notion of a computational turn, has been called *computational philosophy*. Computational philosophy regards the computer as "a medium in which to model philosophical theories and positions" [Bynum and Moor, 1998, p. 6] that can serve as a useful addition to thought experiments and other traditional philosophical methods. In particular, the exploration of philosophical ideas by means of computers allows us to create vastly more complex and nuanced thought experiments that must be made "in the form of fully explicit models, so detailed and complete that they can be programmed" [Grim, Mar and St. Denis, 1998, p. 10]. In addition to fostering the philosophical virtue of precision, it is usually possible to make (real-time) changes to the model, and thereby "explore consequences of epistemological, biological, or social theories in slightly different environments" [Grim, 2004, p.338]. Thus, computer modeling has successfully been applied to philosophy of biology (see also Section 4.2), economics, philosophy of language, physics and logic.Thagard has also pioneered a computational approach to philosophy of science, arguing that computational models can "illustrate the processes by which scientific theories are constructed and used [and] offers ideas and techniques for representing and using knowledge that surpass ones usually employed by philosophers" [1988, p. 2]. Another area in which computer modeling has been employed is ethics. For instance, Danielson [1998] argues that computational modeling of ethical scenarios can help us keep our theories open to counter-intuitive ideas and serve as checks on consistency. Closely related, computer models have also been used to explore topics in social philosophy, such as prejudice reduction [Grim et al, 2005].

Despite the significant advantages, computational philosophy also has limitations. Importantly, it is limited to those kinds of philosophical problems that lend themselves to computational modeling. Additionally, addressing a problem by means of a computer leads to a very specific way of asking questions and placing focus, which might not be equally helpful in all cases. For instance, theories of social dynamics can most easily be computationally modeled by means of rational choice theory, due to its formal nature, which in itself contains particular assumptions that could influence the results (such as methodological individualism). Another problem is that computational modeling can in some cases run counter to fundamental philosophical ideals, because computational models are often built upon earlier computational models or libraries of pre-programmed constructs and, as such, a number of unexamined assumptions can go into a computational model (cf. Grim [2004:339-340]). There are hence reasons for caution in the performance of a computational turn in philosophy. As a matter of fact, the impact of computational modeling on philosophy is as of yet quite limited.

Nevertheless, the notion of a computational turn is referred to explicitly in the mission

statement of the International Association of Computing and Philosophy (IACAP). IACAP, a leading academic organization in the field, was founded in 2004. It was preceded by a conference series in computing and philosophy that started in 1986. In its mission statement, it emphasizes that it does not just aim to promote the study of philosophical issues in computing and IT, but also the use of computers for philosophy. IACAP hence defines a field of "computing and philosophy" that encompasses any interaction between philosophy and computing, including both the philosophy of computing and IT, as defined earlier, and computational philosophy.

In spite of this significant philosophical interest in computer systems, artificial intelligence, and computational modeling, philosophers for a long time paid surprisingly little attention to the very field that made computing possible: computer science. It was not until the late 1990s that philosophers started to pay serious attention to computer science itself, and to develop a true philosophy of computer science. The *philosophy of computer science* can be defined, in analogy with the philosophy of physics or the philosophy of biology, as the philosophical study of the aims, methods and assumptions of computer science. Defined in this way, it is a branch of the philosophy of science. Work in the philosophy of computing did not, or hardly, address questions about the nature of computer science, and the philosophy of AI limited itself to the nature and methods of only one field of computer science, AI.

The relative neglect of computer science by philosophers can perhaps be explained in part by the fact that the philosophers of science has tended to ignore applied science and engineering. The philosophy of science has consistently focused on sciences that aim to represent reality, not on fields that model and design artifacts. With its aim to investigate the nature of intelligence, AI was the only field in computer science with a pretense to represent reality, which may account for much of the attention it received. Other fields of computer science were more oriented towards engineering. In addition, computer science did not have many developed methodologies that could be studied. Methodology had never been the strongest point in such fields as software engineering and information systems. Yet, since the late 1999s, there has been a trickle of studies that do explicitly address issues in computer science [Longo, 1999; Colburn, 2000; Rapaport, 2005; Turner and Eden, 2007a, b], and even an entry in the Stanford Encyclopedia of Philosophy [Turner and Eden, forthcoming b]. The philosophy of computer science is shaping up as a field that includes issues in the philosophy of computing, but that also addresses philosophical questions regarding the aims, concepts, methods and practices of computer science. In section 3, we use the limited amount of literature in this area to lay out a set of issues and problems for the field.

The rise of the personal computer and multimedia technology in the 1980s and the Internet and World Wide Web in the 1990s ushered in a new era in which the computer became part of everyday life. This has brought along major changes in society, including changes in the way people work, learn, recreate and interact with each other, and in the functioning of

organizations and social and political institutions. It has even been claimed that these technologies are fundamentally changing human cognition and experience. These social and cultural changes have prompted philosophers to reflect on different aspects of the new constellation, ranging from the epistemology of hyperlinks to the ontology of virtual environments and the value of computer-mediated friendships. We tie these different investigations together under the rubric *philosophy of the Internet and new media*. Whereas most work in other areas discussed here has been in the analytic tradition in philosophy, a large part of the research in this area is taking place in the Continental tradition, and includes phenomenological, poststructuralist and postmodernist approaches. Additionally, philosophical work in this area is often affiliated with work in social theory and cultural studies. Where appropriate, major works in these areas will be referenced in our survey.

Computer ethics, a fifth area to be surveyed, started out in the late 1970s and gained traction in the mid-1990s, quickly establishing itself as a field with its own journals and conference series. Computer ethics developed largely separately from other areas in the philosophy of computing and IT. Its emergence was driven by concerns of both computer scientists and philosophers about social and ethical issues relating to computers and to address issues of professional responsibility for computer professionals. While its initial emphasis was on professional ethics, it has since broadened to include ethical issues in the use and regulation of information technology in society.

2. Philosophy of Computing

Philosophy of computing is the investigation of the basic nature and principles of computers and the process of computation. Although the term is often used to denote any philosophical issue related to computers, we have chosen to narrow this section to issues focusing specifically on the nature, possibilities and limits of computation. In this section, we will begin by giving an outline of what a computer is, focusing primarily on the abstract notion of computation developed by Turing. We will then consider what it means for something to be computable, outline some of the problems that cannot be computed, and discuss forms of computation that go beyond Turing. Having considered which kinds of problems are Turing non-computable in principle, we then consider problems that are so complex that they cannot be solved in practice. Finally, computing is always computing of something; hence we will conclude this section with a brief outline of central notions like data, representation and information. Since these issues constitute the basics of computing, the many philosophical issues are raised in different contexts and surface in one way or another in most of the following sections. We have chosen to primarily address these issues in the contexts in which they are most commonly raised. In particular, computer science is addressed in section 3, the limits of computation are further addressed in section 4 on artificial

intelligence, and many of the issues regarding computers as (networked) information technologies are discussed in section 5.

2.1 Computation, Computational Systems, and Turing Machines

At the most fundamental level, philosophy of computing investigates the nature of computing itself. In spite of the profound influence computational systems have had in most areas of life, it is notoriously difficult to define terms like 'computer' and 'computation'. At its most basic level, a computer is a machine that can process information in accordance with lists of instructions. However, among many other variations, the information can be analogue or digital, the processing can be done sequentially or in parallel, and the instructions (or, the program) can be more or less sensitive to non-deterministic variables such as user input (see also 2.2 and 4.3). Furthermore, questions regarding computation are sometimes framed in normative terms, e.g. whether it should be defined so as to include the human brain (see Section 4) or the universe at large (see e.g. Fredkin [2003]). At the same time, claims to the effect that computers have had a profound influence on modern society presuppose that there is a distinctive class of artifacts that are computers proper. Indeed, the work of Alan Turing pioneered this development and his notion of a Turing Machine is often invoked in order to explain what computation entails.

Turing's way of defining computation, in effect, was to give an abstract description of the simplest possible device that could perform any computation that could be performed by a human computer, which has come to be known as a Turing machine [Turing, 1937]. A Turing machine is characterized as "a finite-state machine associated with an external storage or memory medium" [Minsky, 1967, p. 117]. It has a read/write head that can move left and right along an (infinite) tape that is divided into cells, each capable of bearing a symbol (typically, some representation of '0' and '1'). Furthermore, the machine has a finite number of transition functions that determines whether the read/write head erases or writes a '0' or a '1' to the cell, and whether the head moves to the left or right along the tape. In addition to these operations, the machine can change its internal state, which allows it to remember some of the symbols it has seen previously. The instructions, then, are of the form, "if the machine is in state a and reads a '0' then it stays in state a and writes a '1' and moves one square to the right". Turing then defined and proved the existence of one such machine that can be made to do the work of all: a Universal Turing Machine (UTM). Von Neumann subsequently proposed his architecture for a computer that can implement such a machine – an architecture that underlies computers to this day.

The purely abstract definition of 'computation' raises a number of controversial philosophical and mathematical problems regarding the in-principle possibility of solving problems by computational means (2.2) and the in-practice possibility of computing highly complex algorithms (2.3). However, it is still debatable whether UTMs really can perform any task that any computer, including humans, can do (see Sections 2.2 and 4). Sloman [2002] and others have

argued that computation, understood in the abstract syntactic terms of a Turing machine or lambda calculus, are simply too far removed from the embodied, interactive, physically implemented and semantic forms of computation at work in both real-world computers and minds [Scheutz, 2002, p. x]. That is, although computation understood in terms of a Turing machine can yield insights about logic and mathematics, it is entirely irrelevant to the way computers are used today – especially in AI research.

2.2 Computability and the Church-Turing Thesis

Computability refers to the possibility of solving a mathematical problem by means of a computer, which can either be a technological device or a human being. The discussion surrounding computability in mathematics had partly been fuelled by the challenge put forward by mathematician David Hilbert to find a procedure by which one can decide in a finite number of operations whether a given first-order logical expression is generally valid or satisfiable [Hilbert and Ackermann, 1928, pp. 73-74; cf. Mahoney, 2004, p. 215]. The challenge to find such a procedure, known as the *Entscheidungsproblem*, led to extensive research and discussion. However, in the 1930's, Church and Turing independently proved that the *Entscheidungsproblem* is unsolvable; Church in terms of lambda calculus and Turing in terms of computable functions on a Turing machine (which were also shown to be equivalent).

In part due to the seminal work of Church and Turing, effectiveness has become a condition for computability. A method is judged to be effective if it is made up of a finite number of exact instructions that requires no insight or ingenuity on the part of the computer and can be carried out by a human being with only paper and pencil as tools. In addition, when such a method is carried out it should lead to the desired result in a finite number of steps. The Universal Turing Machine (UTM) featured prominently in the work of Turing and also in the resulting Church-Turing thesis which holds that a UTM is able to perform any calculation that any human computer can carry out (but see Shagrir [2002] for a distinction between the human, the machine and the physical version of the thesis). An equivalent way of stating the thesis is that any effectively computable function can be carried out by the UTM. On the basis of the Church Turing thesis it became possible to establish whether an effective method existed for a certain mathematical task by showing that a Turing Machine Program could or could not be written for such a task. The thesis backed by ample evidence soon became a standard for discussing effective methods

The development of the concept of the Universal Turing Machine and the Church Turing thesis made it possible to identify problems that cannot be solved by Turing Machines. One famous example, and one of Turing's answers to the *Entscheidungsproblem*, is known as the halting problem. This involves deciding whether any arbitrarily chosen Turing machine will at some point halt, given a description of the program and its input. Sometimes the machine's table

of instructions might provide insight, but this is often not the case. In these cases one might propose to watch the machine run to determine whether it stops at some point. However what conclusion can we draw when the machine is running for a day, a week or even a month? There is no certainty that it will not stop in the future. Similar to the halting problem is the printing problem where the challenge is to determine whether a machine will at some point print '0'. Turing argued that if a Turing machine would be able to tell for any statement whether it is provable through first-order predicate calculus, then it would also be able to tell whether an arbitrarily chosen Turing machine ever prints '0'. By showing that first-order predicate calculus is equivalent to the printing problem, Turing was able to transfer the undecidability result for the latter to the former [Galton, 2005, p. 94]. Additionally, Turing argued that numbers could be considered computable if they could be written by a Turing machine. However since there are only countably many different Turing-machine programs, there are also only countably many computable numbers. Since there are uncountably many real numbers, not all real numbers are computable simply because there are not enough Turing machines to compute them [Barker-Plummer, 2007].

Rice's theorem [Rice, 1953] goes even further and states that there is no algorithm that can decide any non-trivial property of computations [Harel, 2000, p. 54]. More precisely, any non-trivial property of the language recognized by a Turing machine is undecidable. Thus, it is important to recognize that the undecidability problems outlined above, and many more, are not of mere theoretical interest. Undecidability is not an exception, it is the *rule* when it comes to algorithmic reasoning about computer programs (cf. Harel and Feldman [2004]; Harel [2000]).

As these examples show, the Turing machine and the Church-Turing thesis are powerful constructs and can provide deep insights into the nature of computation as well as notions well beyond philosophy of computing. Indeed, Copeland [2004] has argued that some have taken it too far, pointing out many misunderstandings and unsupported claims surrounding the thesis. In particular, many have committed the "Church-Turing fallacy" by claiming that any mechanical model, including the human brain, must necessarily be Turing-equivalent and therefore in-principle possible to simulate on a Turing machine [Copeland, 2004, p. 13]. This claim, sometimes distinguished as the *strong* Church-Turing thesis, presupposes that *anything* that can be calculated by *any* machine is Turing computable, which is a much stronger claim than the thesis that any *effective* method (one that could in-principle be carried out by an unaided human) is Turing computable.

Although Turing proved that problems like the halting problem are unsolvable on any Turing machine, alternative forms of computation have been proposed that could go beyond the limits of Turing-computability – so-called hypercomputation. On a theoretical level, Penrose [1994] has created much controversy by arguing that the human brain is a kind of computer that is capable of mathematical insight unsolvable by a UTM, suggesting that quantum gravity effects are necessary. However, to what degree quantum computers can go beyond UTM, if even

technologically feasible at a grand scale, remains questionable [cf. Hagar, 2007]. MacLennan [2003] has argued that although Turing-computability is relevant to determining effective computability in logic and mathematics, it is irrelevant when it comes to real-time, continuous computation – such as the kind of natural computation found in nature. He further outlines theoretical work that has shown that certain analogue computers can produce non-Turing computable solutions and solve problems like the halting problem (for a comprehensive overview of the history of hypercomputation and its challenges, see Copeland [2002a]). Questions surrounding hypercomputation are primarily of theoretical importance, however, since there is still substantial disagreement on whether a genuine hypercomputer can actually be realized in the physical world (cf. Shagrir and Pitowsky [2003] and Copeland and Shagrir [2007]). The question is also closely related to pancomputationalism and the question whether the universe itself is (hyper-) computational in nature (see e.g. Lloyd [2006] and Dodig-Crnkovic [2006]).

2.3 Computational Complexity

Even in cases where it is in-principle possible to compute a given function, there still remains a question whether it is possible in practice. Theories of computational complexity are concerned with the actual resources a computer requires to solve certain problems, the most central resources being time (or the number of operations required in the computation) and space (the amount of memory used in the computation). One reason why complexity is important is that it helps us identify problems that are theoretically solvable but practically unsolvable. Urquhart [2004] argues that complexity is important to philosophy in general as well, because many philosophical thought experiments do depend on computational resources for their feasibility. If we do take complexity into account, it becomes possible to differentiate between constructs that only exist in a purely mathematical sense and ones that can actually be physically constructed – which in turn can determine the validity of the thought experiment.

Computational complexity theory has shown that the set of problems that are solvable fall into different complexity classes. Most fundamentally, a problem can be considered efficiently solvable if it requires no more than a polynomial number of steps, even in worst-case scenarios. This class is known as P. To see the difference between efficiently solvable and provably hard problems, consider the difference between an algorithm that requires a polynomial (e.g. n^2) and one that requires an exponential (e.g. 2^n) number of operations. If n=100, the former amounts to 10.000 steps whereas the latter amounts to a number higher than the number of microseconds elapsed since the Big Bang. Again, the provably hard problems are not exceptions; problems like Chess and complex route planning can only be achieved by simplified shortcuts that often miss the optimal solution (cf. Harel [2000, pp. 59-89]).

Some problems are easily tractable and some have been proven to require resources way beyond the time and space available. Sometimes, however, it remains a mystery whether

there is a tractable solution or not. The class of NP refers to problems where the answer can be verified for correctness in polynomial time – or, in more formal terms, the set of decision problems solvable in polynomial time by a non-deterministic Turing machine. A non-deterministic Turing machine differs from a normal/deterministic Turing machine in that it has several possible actions it might choose when it is in a certain state receiving certain input; With a normal Turing machine there is always only one option. As a result, the time it would take a non-deterministic Turing machine to compute an NP problem would be the number of steps needed in the sequence that leads to the correct answer. That is, the sequences that turn out to be false do not count towards the number of steps needed to solve the problem, as they do in a normal, deterministic machine. Another way of putting it is to say that the answer to an NP problem can be *verified* for correctness in polynomial time, but the answer itself cannot necessarily be computed in polynomial time (on a deterministic machine). The question, then, becomes: If a given NP problem can be solved in polynomial time on such a machine, is it possible to solve it in polynomial time on a deterministic machine as well? This is of particular importance when it comes to so-called NP-complete problems. A problem is NP-complete when it is in NP and all other NP problems can be reduced to it by a transformation computable in polynomial time. Consequently, if it can be shown that any of the NP-complete problems can be solved in polynomial time, then *all* NP problems can; P=NP. Such a proof would have vast implications, but in spite of tremendous effort and the large class of such problems, no such solution has been found. As a result, many believe that P≠NP, and many important problems are thus seen as being intractable. On the positive side, this feature forms the basis of many encryption techniques (cf Harel [2000, pp. 157ff]).

Traditionally, the bulk of complexity theory has gone into the complexity of sequential computation, but parallel computation is getting more and more attention in both theory and practice. Parallel computing faces several additional issues such as the question of the amount of parallel processors required to solve a problem in parallel, as well as questions relating to which steps can be done in parallel and which need to be done sequentially.

2.4 Data, Information and Representation

Although ‘data’ and ‘information’ are among the most basic concepts in computing, there is little agreement on what these concepts refer to, making the investigation of the conceptual nature and basic principles of these terms one of the most fundamental issues in philosophy of computing. In particular, *philosophy of information* has become an interdisciplinary field of study on its own, often seen as going hand in hand with philosophy of computing. The literature on ‘information’ and related concepts, both historically and contemporary, is vast and cannot be done justice to within this scope (Volume 8 [Adriaans and Benthem, forthcoming] in this series is dedicated to philosophy of information. See also Bar-Hillel [1964], Dretske [1981] and Floridi

[2004a; 2004b; 2007]). In short, the fundamental question in this field is “what is the nature of information?” This question is not only itself illuminated by the nature of computation, but the ‘open problems’ (cf. Floridi [2004a]) in philosophy of information often involve the most fundamental problems in computing, many of which are addressed in other sections (see especially 2.2, 3.3, 3.5, 4.4 and 5.2). It should also be pointed out that this is an area in which philosophy of computing not only extends far beyond computational issues, but also closely intersects with communication studies, engineering, biology, physics, mathematics and cognitive science.

Although it is generally agreed that there can be no information without data, the exact relation between the two remains a challenging question. If we restrict ourselves to computation, it can be added that the data that constitute information must somehow be physically implemented. In practice, data is implemented (or encoded) in computers in binary form, i.e. as some representation of 1 or 0 (on or off), referred to as a *bit*. This satisfies the most fundamental definition of a datum, being “a lack of uniformity between two signs” [Floridi, 2004b, p. 43]. Furthermore, a string of these bits can represent, or correspond to, specific instructions or information. For instance, a computer can be given the instruction ‘1011000001100001’ corresponding to a particular operation, and a computer program can interpret the string ‘01100001’ as corresponding to the letter ‘a’. This underlines, however, that when dealing with questions regarding data, information and representation, it is important to emphasize that there are different levels of abstraction. For instance, a *physical object* can be represented by a *word or an image*, which in turn can be represented by a *string of binary digits*, which in turn can be represented by a *series of on/off switches.* Programming the computer and entering data can be done at different abstraction levels, but the instructions and data have to be converted into machine-readable code (see Section 3.2). The level at which we are operating will determine the appropriate notion of ‘representation’, what it entails to be well-formed and meaningful and whether or not the information must be meaningful *to* someone. With large strings of binary data and a comprehensive and consistent standard that determines what the data refer to (e.g. ASCII), the computer can then output information that is meaningful to a human observer.

As can be seen in the remarks above, there are at least three requirements for something to be information, which is known as the General Definition of Information (GDI); It must consist of data, be well-formed, and (potentially) meaningful. It is, however, controversial whether data constituting semantic information can be meaningful “independent of an informee” [Floridi, 2004b, p. 45]. This gives rise to *one* of the issues concerning the nature of information that has been given extraordinary amount of attention from philosophers: the symbol grounding problem [Harnad, 1990]. In short, the problem concerns how meaningless symbols can acquire meaning, and the problem stems from the fact that for humans, the “words in our heads” have *original* intentionality or meaning (they are *about* something) independently of other observers,

whereas words on a page do not have meaning without being observed – their intentionality is *derived.* However, if it is the case that the human brain is a computational system (or Turing-equivalent), especially when seen as instantiating a "language of thought" (Fodor [1975]; cf. Section 4.2), and if the human brain *can* produce original intentionality, then computers must be able to achieve the same, at least in principle. The problem is perhaps best illustrated by Searle's Chinese room argument [Searle, 1980] where a man inside a room receives symbols that are meaningless to him, manipulates the symbols according to formal rules and returns symbols that are meaningless to him. From the outside it seems as if the response would require an understanding of the meaning of the symbols, but in this case the semantic meaning of the symbols has no bearing on the operations carried out; the meaningfulness of the input and output depends solely on the execution of appropriate formal operations. That is, the semantics going in and out of the system merely supervene on the syntactical data that has been manipulated (or so Searle argues). This is not only one of the central issues in the philosophy of AI, it also constitutes one of the challenges involved in making semantically blind computers perform reliable operations. This is for instance the subject of 'computational semantics', where the aim is to accurately and reliably formalize the meaning of natural language. The main challenges are to define data structures that can deal with the ambiguity and context-sensitivity inherent in natural language and to train or program the computer to make reliable inferences based on such formalizations (cf. Blackburn and Bos [2005]).

3. Philosophy of Computer Science

As argued in the introduction, although philosophers have reflected quite extensively on the nature of computers and computing, they have hardly reflected on the nature of computer *science*. A developed philosophy of computer science therefore currently hardly exists. It is the aim of this section to summarize the scarce philosophical literature that does focus on issues concerning the nature of computer science, and to speculate on what a philosophy of computer science might look like. We hypothesize that a philosophy of computer science would, in analogy to the philosophy of science in general, philosophically reflect on the concepts, aims, structure and methodologies of computer science and its various fields. It would engage in at least the following research activities:

1. *Analysis, interpretation and clarification of central concepts in computer science and the relation between them*. What, for example, is a program? What is data? What is a database? What is a computer model? What is a computer network? What is human-computer interaction? What is the relation between software engineering and computer programming? What is the difference between a programming language and a natural

language? These questions would be answered with the tools and methods of philosophy, and would aim at a philosophical rather than a technical understanding of these concepts. The result would be a deeper, more reflective understanding of these concepts, and possibly an analysis of vaguenesses, ambiguities and inconsistencies in the way that these concepts are used in computer science, and suggestions for improvement.

2. *Analysis, clarification and evaluation of aims and key assumptions of computer science and its various subfields and the relations between them.* What, for example, is the aim of software engineering? What is the aim of operating systems design? How do the aims of different subfields relate to each other? Also, how should these aims be evaluated in terms of their feasibility, desirability, or contribution to the overall aims of computer science? On what key assumptions do various subfields of computer science rest, and are these assumptions defensible?
3. *Analysis, clarification and evaluation of the methods and methodologies of computer science and its various subfields*. What, for example, are the main methodologies used in software engineering or human-computer interaction design? How can these methodologies be evaluated in terms of the aims of these various subfields? What are their strengths and weaknesses? What better methodologies might be possible?
4. *Analysis of the scientific status of computer science and its relation to other academic fields.* Is computer science a mature science or is it still in a preparadigmatic stage? Is computer science a science at all? Is it an engineering discipline? In addition, how do the methodologies of computer science compare to the methods used in natural science, computer science or other scientific fields? Where do the aims of computer science overlap with the aims of other fields, and how do and should computer science either make use of or contribute to other fields? What, for example, is the proper relation between computer science and mathematics, or computer science and logic?
5. *Analysis of the role and meaning of computer science for society as a whole, as well as for particular human aims and enterprises.* How do the aims of computer science contribute to overall human aims? How are the enterprises and projects of computer science believed to make life or society better, and to what extent do they succeed? To what extent is a reorientation of the aims of computer science necessary?

In this section, we will begin with a discussion of attempts to give a general account of the nature, aims and methods of computer science, its status as a science, and its relation to other academic fields. We will then move to five important subfields of computer science, and discuss their nature, aims, methods, and relation to other subfields, as well as any specific philosophical issues that they raise. The subfields that will be discussed are computer programming and software

engineering, data modeling and ontology, information systems, computer simulation, and human-computer interaction. Some areas, such as the nature of programming languages, will naturally be dispersed across many of these sub fields. Another subfield of computer science, artificial intelligence, will be discussed in a separate section because it has generated a very large amount of philosophical literature.

3.1 Computer Science: Its Nature, Scope and Methods

One of the fundamental questions for a philosophy of computer science concerns the nature and scientific status of computer science. Is computer science a genuine science? If so, what is it the science of? What are its distinctive methods, what are its overarching assumptions, and what is its overall goal? We will discuss four prominent accounts of computer science as an academic field and discuss some of their limitations. The first account that is sometimes given may be called the *deflationary* account. It holds that computer science is such a diverse field that no unified definition can be given that would underscore its status as a science or even as a coherent academic field. Paul Graham [2004], for example, has claimed that "computer science is a grab bag of tenuously related areas thrown together by an accident of history", and Paul Abrahams has claimed that "computer science is that which is taught in computer science departments" [Abrahams, 1987, p.1].

An objection to deflationary accounts is that they do not explain how computer science is capable of functioning as a recognized academic field, nor do they address its scientific or academic credentials. Rejecting a deflationary account, others have attempted to characterize computer science as either a science, a form of engineering, or a branch of mathematics [Wegner, 1976; Eden, 2007]. On the *mathematical conception* of computer science, computer science is a branch of mathematics, its methods are aprioristic and deductive, and its aims are to develop useful algorithms and to realize these in computer programs. Theoretical computer science is defended as the core of the field of computer science. A mathematical conception has been defended, amongst others, by Knuth [1974a], who defines computer science as the study of algorithms. Knuth claims that computer science centrally consists of the writing and evaluation of programs, and that computer programs are mere representations of algorithms that can be realized in computers. Knuth defines an algorithm as a "precisely-defined sequence of rules telling how to produce specified output information from given input information in a finite number of steps" [Knuth, 1974a, p.2]. Since algorithms are mathematical expressions, Knuth argues, it follows that computer science is a branch of applied mathematics. A similar position is taken by Hoare [1986, p. 15], who claims: "Computer science is a branch of mathematics, writing programs is a mathematical activity, and deductive reasoning is the only accepted method of investigating programs." The mathematical conception has lost many of its proponents in recent decades, as the increased complexity of software systems seems to make a deductive approach unfeasible.

The *scientific conception* of computer science holds that the apriorism of the mathematical conception is incorrect, and that computer science is an ordinary empirical science. The aim of computer science is to explain, model, understand and predict the behavior of computer programs, and its methods include deduction and empirical validation. This conception has been defended by Allen Newell and Herbert Simon, who have defined computer science as "the study of the phenomena surrounding computers" and who have claimed that it is a branch of natural (empirical) sciences, on a par with ''astronomy, economics, and geology'' [1976, pp. 113-114]. A computer is both software and hardware, both algorithm and machinery. Indeed, it is inherently difficult to make a distinction between the two [Suber, 1988]. The workings of computers are therefore complex causal-physical processes that can be studied experimentally like ordinary physical phenomena. Computer science studies the execution of programs, and does so by developing hypotheses and engaging in empirical inquiry to verify them. Eden claims that the scientific conception seems to make a good fit with various branches of computer science that involve scientific experiments, including "artificial intelligence, machine learning, evolutionary programming, artificial neural networks, artificial life, robotics, and modern formal methods." [2007, p. 138]

An objection to the scientific conception has been raised by Mahoney [2002], who argues that computers and programs cannot be the subject of scientific phenomena because they are not natural phenomena. They are human-made artifacts, and science does not study artifacts but natural phenomena, Mahoney claims. Newell and Simon have anticipated this objection in their 1976 paper, where they acknowledge that programs are indeed contingent artefacts. However, they maintain that they are nonetheless appropriate subjects for scientific experiments, albeit of a novel sort. They argue that computers, although artificial, are part of the physical world and can be experimentally studied just like natural parts of the world (see also Simon [1996]). Eden [2007] adds that analytical methods fall short in the study of many programs, and that the properties of such programs can only be properly understood using experimental methods.

The *engineering conception* of computer science, finally, conceives of computer science as a branch of engineering concerned with the development of computer systems and software that meet relevant design specifications (see e.g. Loui [1987]). The methodology of computer science is an engineering methodology for the design and testing of computer systems. On this conception, computer science should orient itself towards the methods and concepts of engineering. Theoretical computer science does not constitute the core of the field and has only limited applicability. The engineering conception is supported by the fact that most computer scientists do not conduct experiments but are rather involved in the design and testing of computer systems and software. The testing that is involved is usually not aimed at validating scientific hypotheses, but rather at establishing the reliability of the systems that is being developed and in making further improvements in its design.

Eden [2007] has argued that the engineering conception of computer science seems to have won out in recent decades, both in theory and in practice. The mathematical conception has difficulties accounting for complex software systems, and the scientific conception does not make a good fit with the contemporary emphasis on design. A worrisome aspect of this development, Eden argues, is that the field seems to have developed an anti-theoretical and even anti-scientific attitude. Theoretical computer science is regarded to be of little value, and students are not taught basic science and the development of a scientific attitude. The danger is that computer science students are only taught to build short-lived technologies for short-term commercial gain. Eden argues that computer science has gone too far in jettisoning theoretical computer science and scientific approaches, and that the standards of the field has suffered, resulting in the development of poorly designed and unreliable computer systems and software. He claims that more established engineering fields have a strong mathematical and scientific basis, which constitute a substantial part of their success. For computer science (and especially software engineering) to mature as a field, Eden argues, it should embrace again theoretical computer science and scientific methods and incorporate them into methods for design and testing.

3.2 Computer Programming and Software Engineering

Two central fields of computer science are software engineering and programming languages. Software engineering is the "application of a systematic, disciplined, quantifiable approach to the development, operation, and maintenance of software" [Abran et al, 2004]. Theories of programming languages studies the properties of formal languages for expressing algorithms and methods of compiling and interpreting computer programs. Computer programming is the process of writing, testing, debugging and maintaining the source code of computer programs. Programming (or implementation) is an in important phase in the software development process as studied in software engineering.

In theorizing about the nature of software engineering, Parnas has argued that it ought to be radically differentiated from both computer science and programming, and that it should be more closely modeled on traditional forms of engineering. That is, an engineer is traditionally regarded as one "who is held responsible for producing products that are fit for use" [Parnas, 1998, p. 3], which means that software engineering involves a lot more than computer programming and the *creation* of software. Thus, a software engineer should be able to determine the requirements that must be satisfied by the software, participate in the overall design specification of the product, verify and validate that the software meets the requirements, and take responsibility for the product's usability, safety and reliability [Parnas, 1998, p. 3-5]. A similar view of software engineering and its requirements is held by the IEEE Computer Society Professional Practices Committee [Abran et al, 2004]. Software engineering differs, however, from more traditional forms of engineering because software engineers are often unable to avail

themselves of pre-fabricated components and because there is a lack of quantitative techniques for measuring the properties of software. For instance, the cost of a project often correlates with its complexity, which is notoriously difficult to measure when it comes to software [Brookshear, 2007, pp. 326ff]

The importance of software engineering is due to the staggering complexity of many software products, as well as the intricate and often incompatible demands of shareholders, workers, clients and society at large. This complexity is usually not the kind of computational complexity described in 2.3, but rather the complexity involved in specifying requirements, developing a design overview, as well as verifying and validating that the software satisfies internal and external requirements. The verification and validation of software is a critical part of software engineering. A product can work flawlessly but fail to meet the requirements set out initially, in which case it fails *validation* ("The right product was not built"). Or, it can generally meet the requirements set out initially, but malfunction in important ways, in which case it fails *verification* ("The product was not built right"). The methods employed in verification often reflect the overall perspective on what computer science and computer programs are. Eden [2007] outlines three paradigms of computer science (cf. Section 3.1), in which software is verified by means of a priori deductive reasoning (rationalist), by means of a posteriori, empirical testing (technocratic) or by means of a combination (scientific). The rationalist paradigm is most closely related with the question of 'formal program verification'. This long-lasting debate is concerned with the question whether software reliability can (in some cases) be ensured by utilizing deductive logic and pure mathematics [Fetzer, 1988; 1991; 1998]. This research stems from dissatisfaction with "technocratic" means of verification, including manual testing and prototyping, which are subjective and usually cannot guarantee that the software is reliable (cf. Section 2.2). Clearly, proponents of formal program verification tend to regard computer science as analogous to mathematics. This perspective is especially evident in Hoare [1986] who regards computers as mathematical machines, programs as mathematical expressions, programming languages as mathematical theories and programming as a mathematical activity. However, as mentioned in 3.1, the complexity involved in modern software engineering has left the mathematical approach unfeasible in practice.

Although software engineering encompasses a range of techniques and procedures throughout the software development process, computer programming is one of the most important elements. Due to the complexity of modern software, hardly anyone programs computers in machine code anymore. Instead, programming languages (PL) at a higher abstraction level are used, usually being closer to natural language constructs. These source codes are then *compiled* into instructions that can be executed by the computer. Indeed, programming languages can be seen as 'virtual machines', i.e. abstract machines that do not exist, but must be capable of being translated into the operations of an existing machine

[McLaughlin, 2004, p. 139]. Based on these principles, thousands of different programming languages have been developed, ranging from highly specialized and problem-specific to multi-purpose, industry-standard PLs.

The investigation of appropriate mechanisms for abstraction is one of the primary concerns in the design and creation of both computer programs and the programming languages themselves [Turner and Eden, forthcoming b]. The characteristics of program abstraction can be explained through Quine's distinction between choosing a scientific theory and the ontological commitments that follows. That is, whereas the choice of a scientific theory is a matter of explanatory power, simplicity and so forth, once a theory has been selected, existence is determined by what the theory says exists [Turner and Eden forthcoming a, p. 148; Quine, 1961]. In computer science terms, once the choice of PL has been made, the PL more or less forces the programmer to solve problems in a particular way – within a given conceptual framework. This underlying conceptual framework can be referred to as the programming *paradigm.* The initial choice of programming language (or paradigm) depends on a number of factors, primarily its suitability for the problem at hand.

However, the notion of programming *paradigm* carries some of the more irrational connotations of Kuhn's [1970] concept, meaning that the use of a particular PL is often determined by social, commercial and ad hoc considerations, and sometimes lead to polarization and lack of communication within the field of software engineering [Floyd, 1978]. These ontological commitments do concern questions considered with regard to data modeling (see Section 3.3), but are more closely related to the control structures that operate on the data. For instance, abstraction necessarily entails some form of 'information hiding'. This is, however, a different kind of abstraction than that found in formal sciences. In many sciences, certain kinds of information are deemed irrelevant, such as the color of triangle in mathematics, and therefore neglected. In PL abstraction, as Colburn and Shute [2007] has pointed out, information that is "hidden" at one level of abstraction (in particular, the actual machine code needed to perform the operations) cannot be ignored at a lower level.

Another example concerns constructs that are used as high-level shorthands, but whose exact nature might not be preserved when compiled into machine-readable code, such as random-number generators that can only be quasi-random or fractions computed as truncated decimals. Finally, PLs differ immensely with regard to the structure and flow of the control structures. For instance, Edsger Dijkstra's seminal paper "Go To Statement Considered Harmful" [Dijkstra, 1968], which has spurred dozens of other "x considered harmful" papers, criticized the then common use of unstructured jumps (goto's) in programming, advocating a structured approach instead. Interestingly, discussions surrounding 'good' and 'bad' programming differ enormously depending on the underlying justification, whether it is ease of learning, reliability,

ease of debugging, ease of cooperation or, indeed, a notion of aesthetic beauty (see e.g. Knuth [1974b]).

3.3 Data Modeling and Ontology

One of the most common uses of computer technology, and a central concern in computer programming and software engineering, is to store vast amounts of data in a database so as to make the storage, retrieval and manipulation as efficient and reliable as possible. This requires a specification beforehand of how the database should be organized. Such a specification is known as a data model theory. Although the term is used in many different senses (cf. Marcos [2001]), a data model typically consists of 1) the structural part, i.e. a specification of how to represent the entities or objects to be modeled by the database; 2) the integrity part, i.e. rules that place constraints in order to ensure integrity; and 3) the manipulation part, i.e. a specification of the operations that can be performed on the data structures. The purpose of these parts is to ensure that the data are stored in a consistent manner, that queries and manipulations are reliable and that the database preserves its integrity. Data integrity refers to the accuracy, correctness and validity of the data, which in lack of a comprehensive data model theory, might be compromised when new data is entered, when databases are merged or when operations are carried out. To ensure integrity in human interactions with the database, such interaction is usually regulated by a Database Management System. Furthermore, we can distinguish between 2-dimensional databases, which can be visualized as a familiar spreadsheet of rows and columns, and n-dimensional databases, where numerous databases are related to each other, for instance by means of shared 'keys'.

Floridi makes a distinction between an 'aesthetic' and a 'constructionist' view concerning the nature and utility of databases (Floridi [1999]; see Marcos and Marcos [2001] for a similar distinction between 'model-as-copy' and 'model-as-original'). First, the "aesthetic" approach sees databases as a collection of data, information or knowledge that conceptualizes a particular reality, typically modeled on naïve realism. This approach can in particular be seen in 'knowledge engineering', where human knowledge is collected and organized in a 'knowledge base', usually forming the basis of an 'expert system' (see Section 4.4). In a similar vein, Gruber [1995] defines the use of *ontology* in computer science as "a specification of a representational vocabulary for a shared domain of discourse [including] definitions of classes, relations, functions, and other objects" [Gruber, 1995, p. 908]. Although this is the most common use of data modeling, one of the philosophical problems with such "specification of conceptualization" is that these conceptualizations might not directly correspond to entities that exist in the real word but to human-constructed concepts instead. For instance, these conceptualizations have much in common with folk psychological concepts, whose validity has been contested by many

philosophers (cf. Barker [2002]). This is particularly a problem when it comes to science-based ontologies, where non-existing entities ought to be avoided [Smith, 2004]. According to Floridi, a second approach to data modeling can be termed 'constructionist', where databases are seen as a strategic resource whose "overall purpose is to generate new information out of old data" and the implemented data model "is an essential element that contributes to the proper modification and improvement of the conceptualized reality in question" [Floridi, 1999, p. 110). The distinction between 'aesthetic' and 'constructionist' also gives rise to an epistemological distinction between those sciences where the database is intended to represent actual entities, such as biology and physics, and those sciences where databases can provide requirements that the implementation in the real world must satisfy, including computer science itself [Floridi, 1999, p. 111].

Although data model theories are application- and hardware-independent, they are usually task-specific and implementation-oriented. This has raised the need for domain- and application-independent ontologies*,* the purpose being to establish a high-level conceptualization that can be shared by different data models – in different domains. Since such ontologies often aim to be task-independent, they typically describe a hierarchy of concepts, properties and their relations, rather than the entities themselves. This is known as a *formal,* as opposed to a *descriptive* ontology and is influenced by philosophical attempts to develop ontological categories in a systematic and coherent manner. The impetus of much of this research stems from a common problem in computer science, sometimes referred to as the tower of Babel problem. Especially with the advent of networked computers, the many different kinds of terminals, operating systems and database models – as well as the many different domains that can be represented in a database – posed a problem for successful exchange of data. Rather than dealing with these problems on an *ad hoc* basis, formal ontologies can provide a common controlled vocabulary [Smith et al, 2007, p. 1251] that ensures compatibility across different systems and different types of information. Such compatibility does not only save man hours, but opens up new possibilities for cross-correlating and finding "hidden" information in and between databases (so-called 'data mining'). The importance of such ontologies has been recognized in fields as diverse as Artificial Intelligence and knowledge engineering (cf. Section 4), information systems (cf. 3.4), natural language translation, mechanical engineering, electronic commerce, geographic information systems, legal information systems and, with particular success, biomedicine (cf. Guarino [1998]; Smith et al [2007]). Paradoxically, however, the very success of this approach has led to a proliferation of different ontologies that sometimes stand in the way of successful integration [Smith et al, 2007]. Closely related, these ontologies cannot always cope with specific domain-dependent requirements. This could be one reason why, despite the philosophical interest and heritage, the importance of philosophical scrutiny have often been "obscured by the temptation to seek immediate solutions to apparently localized problems" [Fielding et al, 2004]. This tension especially arises in software engineering, in which the

theoretical soundness of the ontology must often be balanced by real-world constraints (see Section 3.2)

3.4 Information systems

'Information' and 'system' are both highly generic terms, which means that the term 'information system' is used in many different ways. In light of this, Cope et al [1997] conducted a survey of different uses of the term and identified four major conceptions that form a hierarchy. At one end, the most general conception of IS simply refers to a database where information can be retrieved through an interface. At the other end, the more specific conception considers IS to encompass the total information flow of a system, typically a large organization – including "people, data, processes, and information technology that interact to collect, process, store, and provide as output the information needed to support an organization" [Whitten, 2004, p. 12]. As such, IS is not the same as information technology but a *system* in which information technology plays an important role. Du Plooy also argues that the social aspect of information systems is of such importance that it should be seen as the core of the discipline [du Plooy, 2003] and we will focus on this notion in this sub section, given that many of the non-social issues are discussed elsewhere

Although IS includes many factors in addition to the technology, the focus in IS research has typically been on the role of the technology, for instance how the *technology* can be optimized to improve the information flow in an organization. Among the many philosophical issues raised by such systems, one of the most important ones are the relation between scientifically-based, rationalist theories of information systems design and the actual practice of people involved in management. Introna [1997] argues that the (then) reigning techno-functionalist paradigm in the information systems discipline fails to take actual practices into account. Based on insights from hermeneutics, he stresses instead the importance of the involved manager and the changing demands of being-there as a part of the information system. In a similar manner, Butler and Murphy [2007] argue that computerization of organizations means that we rationalize what is easy to rationalize, and therefore place too much emphasis on decontextualized information processes rather than the reality of the human actors. As can be seen in these examples, theories of information systems often address the (power) relationship between humans and technology – especially the over-emphasis on technology at the expense of humans – which means that hermeneutics and theorists like Giddens, Heidegger, Habermas, Foucault and Latour often lend themselves to such analysis.

It should also be pointed out that IS research often involves assessment of actual information systems and as such pre-supposes certain methodologies and assessment criteria. Dobson points out that this raises a number of epistemological questions regarding the IS researcher's theoretical lens, skill and political biases, as well as a number of ontological

questions regarding which entities to include as part of the information system and their relation to each other [Dobson, 2002]. Given the complexity of large information systems, the ontological and epistemological questions become particularly challenging, because large organizations often involve fundamentally different kinds of entities. For instance, an Information system can include specifications of such things as databases of physical entities, algorithms for efficient scheduling, information flow relations, and interfaces for retrieving information (see Section 3.3 for more on computer ontologies). Dobson argues that the adopted methodologies have been dominated by the kind of social theorists mentioned above. He suggests that IS studies should pay more attention to philosophical approaches to epistemology and ontology, and further suggests that Bhaskar's critical realism is one important approach because it sees philosophy as operating at the same level as methodological issues, and because of its acknowledgment that observation is value-laden, i.e. the choice and assessment of information systems is partly determined by political biases and values concerning organization hierarchies, workers' rights and so forth (cf. Dobson [2002]).

Many of the perennial issues in computer ethics also revolve around information systems, including problems surrounding surveillance in the workplace, automation of manual labor and the problems with assigning responsibility (see also Sections 4.6 and 6). These are issues that usually fall under professional ethics because they deal with computer science and information systems professionals – management in particular. Other issues in philosophy of information systems overlap with philosophy of computer science in general (3.1), software engineering (3.2), and often address issues concerned with data modeling and ontology (3.3) as well as human-computer interaction (3.6).

3.5 Computer Simulation

In addition to the kind of data modeling outlined above, which is primarily occupied with commercial and management systems, computers are frequently employed for scientific modeling and simulation. Computers are particularly useful when simulating micro- or macroscopic phenomena, where traditional forms of experimentation is not feasible. Computer simulations also differ from data models in that they often employ visualization of the data (especially simulations at micro- or macro level) or real-time input from users (e.g. flight simulation).

One of the main issues in the establishment of philosophy of computer science as distinct from traditional philosophy of science revolves around the latter's ability to adequately account for computer simulations. Clearly, computers allow us to simulate events that we have not been able to simulate before, primarily because of the immense processing power and visualization possibilities, and therefore becomes a valuable tool for many sciences. The question remains, however, whether computer simulation raises any novel philosophical issues. Humphreys

[forthcoming] argues that computer simulation raises the need for a new philosophy of computer science in a number of ways. Most importantly, computer simulations are epistemically opaque in the sense that it is often impossible for a researcher to know all of the epistemically relevant elements of the process. This is not only a result of the complexity involved, but also the fact that scientists need to delegate substantial amounts of authority to the programmers and software engineers. Furthermore, the question of what we can know in philosophy of science has become in part a question of computational possibilities and limitations. Similar concerns have been raised by Brian Cantwell Smith, who argues that there are inherent limitations to what can be proven about computers and computer programs; the 'correctness' of a computer simulation is vulnerable to the combination of computational complexity, unpredictable human-computer interaction, the many levels at which a computer can fail and the lack of precision regarding what 'correctness' entails [Smith, 1996].

Morgan and Morrison [1999] have argued that it is the lack of material similarity that makes computer simulations unique; that computer simulations thereby have a reduced potential to make strong inference back to the world. Parker [forthcoming] and others have argued against this view by pointing out that non-computer simulations often have less validity because of the complexity involved – for instance with respect to weather forecasting. Winsberg [forthcoming], although disagreeing with Morgan and Morrison, makes a related but ultimately more plausible argument. He argues that it is not the degree of materiality that determines the validity of the simulation in question, but the justification of this validity itself. Thus, the difference between traditional forms of simulation and computer simulation lies in the fact that the validity of non-computer simulations can be justified by pointing to shared material properties whereas computer simulations cannot. In other words, non-computer simulations typically justify the similarity between the object and target systems in terms of material or structural similarities (e.g. a miniature airplane in a wind tunnel being similar to its full-scale counterpart in some material respects), whereas the validity of computer simulations are typically justified in terms of strictly formal and mathematical similarities. In other words, computer simulations require a justification that is entirely different from, and raise philosophical questions typically not raised by, non-computer simulations – such as the relation between formalized algorithms and the physical world, the justifiability of heuristic algorithms, the use of pseudo-randomizations and truncated numbers, approximations of physical laws and so forth. Frigg and Reiss [forthcoming] argue that despite these characteristics, computer simulations pose no new philosophical problems. But, as Humphrey [forthcoming] points out, this conclusion seems to rest on a misinterpretation of the Church-Turing thesis (cf. Section 2.2). That is, although computer simulations in-principle can be carried out on a Turing machine and *ipso facto* by means of pencil and paper, the staggering complexity of e.g. weather forecasting forbids this in practice.

Computer simulations also offer the possibility to visualize the results of the simulation and even to let the researchers intervene on the basis of visual feedback – thereby resembling *experiments* ("*in silico*") more than simulations. In particular, these visualizations can be run at different speeds, in reverse and with different foci. Thus, computer simulation has opened up simulation possibilities that are otherwise prohibited by physical laws. This is particularly the case when virtual reality is used for literally seeing things that are hidden in complex algorithms and insurmountable amounts of raw data.

3.6 Human-Computer Interaction

Human-Computer Interaction (HCI) is a subfield within computer science concerned with the study of the interaction between people (users) and computers and the design, evaluation and implementation of user interfaces for computer systems that are receptive to the user's needs and habits. It is a multidisciplinary field, which incorporates computer science, behavioral sciences, and design. A central objective of HCI is to make computer systems more user-friendly and more usable. Users interact with computer systems through a user interface, which is the hardware and software through which users and computer systems communicate or interact with each other. The user interface provides means of input, allowing users to manipulate the system, and output, allowing the system to provide information to the user. The design, implementation and evaluation of interfaces is therefore a central focus of HCI.

It is recognized in HCI that good interface design presupposes a good theory or model of human-computer interaction, and that such a theory should be based in large part on a theory of human cognition to model the cognitive processes of users interacting with computer systems [Peschl & Stary, 1998]. Such theories of human cognition are usually derived from cognitive psychology or the multidisciplinary field of cognitive science. Whereas philosophers have rarely studied human-computer interaction specifically, they have contributed significantly to theorizing about cognition, including the relation between cognition and the external environment, and this is where philosophy relates to HCI.

Research in HCI has initially relied extensively on classical conceptions of cognition as developed in cognitive psychology and cognitive science. Classical conceptions, alternatively called *cognitivism* or the *information-processing approach*, hold that cognition is an internal mental process that can be analyzed largely independently of the body of the environment, and which involves the manipulation of discrete, internal states (representations or symbols) that are manipulated according to rules or algorithms [Haugeland, 1978]. These internal representations are intended to correspond to structures in the external world, which is conceived of as an objective reality fully independent of the mind. Cognitivism has been influenced by the rationalist tradition in philosophy, from Descartes to Jerry Fodor, which construes the mind as an entity separate from both the body and the world, and cognition as an abstract rational, process. Critics

have assailed cognitivism for these assumptions, and have argued that cognitivism cannot explain cognition as it actually takes place in real-life settings. In its place, they have developed *embodied* and *situated* approaches to cognition that conceive of cognition as a process that cannot be understood without intimate reference to the human body and to the interactions of humans with their physical and social environment [Anderson, 2003]. Many approaches in HCI now embrace an embodied and/or situated perspective on cognition.

Embodied and situated approaches share many assumptions, and often no distinction is made between them. Embodied cognition approaches hold that cognition is a process that cannot be understood without reference to the perceptual and motor capacities of the body and the body's internal milieu, and that many cognitive processes arise out of real-time goal-directed interactions of our bodies with the environment and thus have to consider sensorimotor interactions of the body with the environment as integral to the cognitive process. Situated cognition approaches hold that cognitive processes are co-determined by the local situations in which agents find themselves. Knowledge is constructed out of direct interaction with the environment rather than derived from prior rules and representations in the mind. Cognition and knowledge are therefore radically context-dependent and can only be understood by considering the environment in which cognition takes place and the agent's interactions with this environment. Together, embodied and situated approaches present a conception of cognition as embodied, engaged, situated and constructive, in which an understanding the agent's bodily interactions with an environment are essential to an understanding of cognition.

Embodied and situated approaches have been strongly influenced by phenomenology, especially the work of Heidegger and Merleau-Ponty, and the contemporary work of Hubert Dreyfus. Lucy Suchman, one of the founders of the field of HCI and an early proponent of a situated approach [Suchman, 1987] is even a student of Dreyfus. Many other proponents of embodied/situated approaches in HCI make extensive reference to phenomenology [e.g., Winograd & Flores, 1987; Dourish, 2001].

Philosophers Andy Clark and David Chalmers have developed an influential embodied/situated theory of cognition, *active externalism*, according to which cognition is not a property of individual agents but of agent-environment pairings. They argue that external objects play a significant role in aiding cognitive processes, and that therefore cognitive processes extend to both mind and environment. This implies, they argue, that mind and environment together constitute a cognitive system, and the mind can be conceived of as extending beyond the skull [Clark and Chalmers, 1998; Clark, 1997]. Clark uses the terms "wideware" and "cognitive technology" to denote structures in the environment that are used to extend cognitive processes, and he argues that because we have always extended our minds using cognitive technologies, we have always been cyborgs [Clark, 2003]. Active externalism has been inspired by, and inspires, *distributed cognition* approaches to cognition [Hutchins, 1995], according to which

cognitive processes may be distributed over agents and external environmental structures, as well as over the members of social groups. Distributed cognition approaches have been applied to HCI [Hollan, Hutchins and Kirsh, 2000], and have been especially influential in the area of Computer Supported Cooperative Work (CSCW).

Brey [2005] has invoked cognitive externalist and distributed cognition approaches to analyze how computer systems extend human cognition in human-computer interaction. He claims that humans have always used dedicated artifacts to support cognition, which cognitive scientist and HCI researcher Donald Norman [1993] has called *cognitive artifacts*. These are artifacts designed to represent, store, retrieve or manipulate information. Computer systems are extremely versatile and powerful cognitive artifacts that can support almost any cognitive task. They are capable of engaging in a unique symbiotic relationship with humans to create hybrid cognitive systems in which a human and an artificial processor process information in tandem. However, Brey argues, not all uses of computer systems are cognitive. With the emergence of graphical user interfaces, multimedia and virtual environments, the computer has become a simulation device next to a cognitive device. Computers are now often used to simulate environments to support communication, play, creative expression, and social interaction. Brey argues that while such activities may involve distributed cognition, they are not primarily cognitive themselves. Interface design has to take into account whether the primary aim of applications is cognitive or simulational, and different design criteria exist for both.

4. Philosophy of Artificial Intelligence

Artificial Intelligence (AI) is commonly referred to as the science and engineering of intelligent machines – 'intelligent' commonly seen as relative to *human* intelligence. Given its close ties with numerous sub disciplines of philosophy, philosophy of mind and philosophy of language in particular, it has received tremendous attention from philosophers. The field is inherently interdisciplinary and has arguably had a more profound impact on philosophical discourse than any other technology. This section discusses issues and approaches in the philosophy of artificial intelligence, including its emergence and scope, the philosophy of major approaches in AI (symbolic AI, connectionist AI, artificial life, dynamical systems), the philosophy of AI applications (expert systems, knowledge engineering, robotics, and artificial agents) and concludes with a review of some ethical issues in AI.

4.1 Artificial Intelligence and Philosophy

Artificial intelligence, or AI, is a field of computer science that became established in the 1950s. It was described at the time as a new science which would systematically study the phenomenon of 'intelligence'. This goal was to be pursued by using computers to simulate intelligent processes.

The central assumption of AI was that the logical operations of computers could be structured to imitate human thought processes. Because the workings of a computer are understood while those of the human mind are not, AI researchers hoped in this way to reach a scientific understanding of the phenomenon of 'intelligence'.

Intelligence is conceived of in AI as a general mental ability that encompasses several more specific abilities, such as the ability to reason, plan, solve problems, comprehend ideas, use language, and learn. AI research commonly focuses on a specific ability and attempts to develop programs that are capable of performing limited tasks involving that ability. The highest goal of AI was to construct a computer system with the intelligence and reasoning ability of an adult human being. Many early AI researchers claimed that this goal would be reached within only a few decades, thanks to the invention of the digital computer and to key breakthroughs in the fields of information theory and formal logic. In 1965, the noted AI researcher Herbert Simon predicted that computers would be able to execute any task that human beings could by 1985 [Simon, 1965]. Marvin Minsky, another key figure in AI, predicted in 1967 that all of AI's important goals could be realized within a generation [Minsky, 1967].

How might it be demonstrated that a computer is as intelligent as a human being? Alan Turing proposed that a machine is demonstrably intelligent if it is able to fool human beings into thinking that it may be human [1950]. In the *Turing Test*, a computer and a human being are placed behind a screen, and a test person is to ask questions to both in order to find out which of the two is human. If the test person cannot make such a judgment after a reasonable amount of time, the computer has passed the test and it has supposedly been demonstrated to be in possession of general intelligence. The Turing Test is still often invoked, but has not remained without criticism as a test for general intelligence [Moor, 2003].

AI researchers agreed that AI studied intelligent processes and aimed to create intelligent computer programs, but they soon developed different viewpoints on the extent to which AI should be directed at the study of *human* intelligence. Some researchers, like Allen Newell and Herbert Simon, believed that intelligent computer programs could be used to model thought processes of humans, and made it their goal to do this. This is sometimes called the *cognitive simulation* approach in AI, or *strong AI* [Searle, 1980]. Strong AI holds that suitably programmed computers literally have cognitive states that resemble the cognitive states found in human minds, and are therefore capable of explaining human cognition. Some proponents of strong AI even go further and hold that a suitably programmed computer is capable of consciousness.

Underlying these claims of strong AI is a belief in *computationalism*: the doctrine that mental states are computational states, and that cognition equals computation [Pylyshyn, 1984; Shapiro, 1995]. In the mid-1970s, computationalism became a widely held view within AI, linguistics, philosophy, and psychology, and researchers from these fields joined to create the

field of *cognitive science*, a new field that engages in interdisciplinary studies of the mind and intelligence [Boden, 2006].

Whereas many researchers in AI embraced the cognitive simulation approach, many others merely wanted to develop computer programs that were capable of performing intelligent tasks. For all they were concerned, the underlying mechanism by which computers were capable of intelligent behavior might be completely different from the working of human minds. This approach has been called *weak AI*. Many proponents of this more cautious approach nevertheless believed that research in AI could contribute to an understanding of the phenomenon of intelligence, by uncovering general properties of intelligent processes, and that AI could therefore still meaningfully contribute to cognitive science.

In recent decades, the view of AI as a science that studies the phenomenon of intelligence has been partially superseded by a view of AI as an engineering discipline. Instead of trying to understand intelligence, most contemporary AI researchers focus on developing useful programs and tools that perform in domains that normally require intelligence. AI has therefore in large part become an applied science, often merging with other fields of computer science to integrate AI techniques into areas such as data mining, ontological engineering, computer networking, agent technology, robotics, computer vision, human-computer interaction, ubiquitous computing and embedded systems.

The philosophy of AI [Copeland, 1993; Haugeland, 1981; Boden, 1990; Fetzer, 2004] emerged in the 1960s and became an established field in the 1980s. For the most part, it focuses on assumptions and approaches within the scientific approach to AI, and its relation to cognitive science. Much less attention has been paid to developments in the engineering approach to AI The philosophy of AI considers the questions whether machines (and specifically computer systems) are capable of general intelligence, whether they are capable of having mental states and consciousness, and whether human intelligence and machine intelligence are essentially the same and the mind therefore is a computational system. Philosophers have also explored the relation between philosophical logic and AI [Thomason, 2003] and ethical issues in AI (Section 4.6).

4.2 Symbolic AI

From the beginnings of AI research in the 1950s up to the early 1980s different approaches in AI research had so much in common that they constituted a research paradigm, in the sense articulated by Kuhn. This research paradigm has been called "symbolic AI" (or, alternatively, "classical AI" or GOFAI, which stands for Good Old Fashioned AI) and is still influential today. The central claim of symbolic AI is that intelligence, in both humans and machines, is a matter of manipulating symbols according to fixed and formal rules. This claim rests on several basic assumptions, made precise by Newell and Simon [1976], who introduced the notion of a *physical*

symbol system. A physical symbol system was defined by them as a system that manipulates and produces physically realized symbol structures. Symbol structures, or expressions, are physical combinations of instances of symbols, which are unique physically realized patterns. The system contains continually changing sets of symbol structures as well as a set of processes for their creation, modification, reproduction and destruction. Symbol structures are capable of designating objects in the world and are also capable of designating processes which can be carried out ("interpreted") by the system. Clearly, computer systems qualify as physical symbol systems, but the above definition leaves open the possibility that other entities, such as human brains, are also physical symbol systems.

Based on this definition, Newell and Simon state the *physical symbol system hypothesis*, which is that a physical symbol system has the necessary and sufficient means to display general intelligence. Because this hypothesis implies that only physical symbol systems can display general intelligence, it also implies that the human mind implements a physical symbol system, and that minds are information-processing systems very similar to digital computers. This view is called computationalism, or the *computational theory of mind*. Newell and Simon's hypothesis is therefore a version of strong (symbolic) AI. A weaker version, equivalent to weak (symbolic) AI is that being a physical symbol system is sufficient but not necessary for intelligence, which implies that computer systems are capable of general intelligence but that their architecture may not resemble that of human minds [Copeland, 1993].

Strong symbolic AI, and the corresponding computational theory of mind, have been both defended and criticized by philosophers. Philosopher Jerry Fodor has famously defended computationalism through a defense of his Language of Thought Hypothesis, which states that human cognition consists of syntactic operations over physically realized representations in the mind that have a combinatorial syntax and semantics [Fodor, 1975].

The most famous (or infamous) critic of strong symbolic AI, John Searle, has argued that when computers process symbols they do not have access to their content or meaning, whereas humans do, and that therefore computationalism and strong symbolic AI are false. He makes his case using a thought experiment, called the *Chinese Room Argument*, in which a human in a room is asked to follow English instructions for manipulating Chinese symbols [Searle, 1980]. The human receives questions in Chinese through a slot in the wall, and is capable of answering in Chinese, thus appearing to understand Chinese. Yet, Searle claims, the human does not understand Chinese. The example shows, he argues, that manipulating symbols on the basis of syntax alone, which is what the human does, does not imply understanding. Computer cognition and human cognition are therefore different, because humans normally do have understanding of the information they process. Searle's argument against strong symbolic AI has met with numerous responses and attempted rebuttals from the AI community and fellow philosophers [Preston & Bishop, 2002].

Hubert Dreyfus, who has been critiquing AI since the mid-1960s, has argued like Searle that human cognition is fundamentally different from information processing by computers [Dreyfus, 1972; 1992; 1996]. Dreyfus derives his view of human cognition from phenomenology. Human cognition, he claims, normally does not involve the application of rules, and does not normally make use of internal representations. Dreyfus holds instead that human intelligence is *situated* - codetermined by the situation in which humans find themselves – and *embodied* – emergent out of real-time goal-directed sensorimotor interactions of the human body with the environment. Computers, he argues, are disembodied, and their information processing is not situated but detached and abstracted from the world in which they find themselves. They are therefore fundamentally different from human minds. Since the 1980s, situated and embodied approaches became an important alternative approach in AI research (see also Section 3.6).

The assumptions of symbolic AI about the nature of intelligence are so fundamentally mistaken, Dreyfus argues, that weak symbolic AI is false as well. Symbolic AI is therefore in-principle incapable of yielding general intelligence. The problem, Dreyfus argues, is that symbolic AI stands in the tradition of Cartesian rationalism, and inherits all of its false assumptions: that intelligence involves the disembodied application of formal rules, that the world we know has a formal, objective structure, and that all knowledge can be formalized. Dreyfus is particularly critical of the third assumption, which he calls the *epistemological assumption*. This assumption implies that everything that is known or understood by humans can be expressed in context-independent, formal rules or definitions that can be processed by machines.

Dreyfus argues against this assumption that, while formal rules may be one way of *describing* human knowledge, they do not provide the basis for *reproduction* of such knowledge by an intelligent system. The problem is that formal rules do not contain their own criteria for application, and that additional contextual or background information is needed. The problem with computers is that they do not possess common sense, Dreyfus argues. They do not possess the elaborate system of background information possessed by humans in virtue of which they can interpret items effortlessly in the context in which they occur, and by which they know which interpretations are meaningful and which ones absurd or meaningless. Dreyfus calls this problem for symbolic AI the *commonsense knowledge problem*, and claims that it is unsolvable within a symbolic approach [Dreyfus & Dreyfus, 1986]. Many “hard” problems in symbolic AI, such as the well-known *frame problem* [Pylyshyn, 1987], can be analyzed as specific instances of the commonsense knowledge problem.

4.3 Connectionist AI, Artificial Life and Dynamical Systems

Since the 1980s, a rival paradigm to symbolic AI has arisen, called *neural networks* or *connectionism* [Bechtel and Abrahamson, 1990; Clark, 1991]. Connectionist AI is often viewed as a radical alternative to symbolic AI, rejecting from the start the idea that intelligent behavior

springs from the manipulation of symbols according to formal rules. The neural network approach derives its inspiration for the modeling of intelligent processes from the structure and operation of the human brain rather than from digital computers.

Connectionist models consist of a large number of simple processors, or units, with relatively simple input/output functions that resemble those of nerve cells. These units are connected to each other and some also to input or output structures, via a number of connections. These connections have different “weight”. The weights in combination with the input signals determine the activation level of a unit. Units may be activated to different degrees and when the activation reaches a certain threshold they give off signals to connected units. A complete connectionist network consists of an input layer of input units, an output layer, and one or more “hidden” layers of units in between. Information processing in a connectionist system is then a process of excitation and inhibition of units. It is a massively parallel process, in which large numbers of simple computational units perform simple computations and influence each other, ultimately leading to an set of output signals. Representations in connectionism can be defined as patterns of activation across a unit layer.

Neural networks turn out to be astoundingly good at carrying out certain types of intelligent tasks, like pattern recognition, categorization, and the coordination of behavior. They have been less successful, so far, in modeling “higher” cognitive tasks, like abstract reasoning, formal tasks, and problem solving, precisely the kinds of tasks that symbolic AI is best able to model. Attempts have been made to physically build such connectionist models, but in practice, most connectionist models are simulated on ordinary digital computers.

Connectionist and symbolic AI share the assumption that cognition is a matter of information processing, and that such information processing is computational, meaning that it can be represented algorithmically and mathematically. There are four major differences between the two approaches. First, information processing in symbolic AI involves the application of explicit, formal rules, whereas no such rules are operative in connectionist networks. Connectionist processors are computing units that do not act on the syntactic properties of input signals but merely to their strength. Second, information processing in symbolic AI is executive-driven, involving a central overviewer (processing unit) which controls processes, whereas information processing in networks is the result of many independently operating structures. Third, information processing in symbol systems is typically serial, whereas in networks it is massively parallel. Fourth, learning in symbolical AI is a deductive process consisting of hypothesis testing, whereas in connectionism, it is associationist, i.e. a process of strengthening or weakening (or growing or losing) connections between nodes.

Connectionism was embraced enthusiastically by many philosophers in the 1980s and 1990s as a superior alternative to symbolic AI. Amongst its strongest proponents were Andy Clark [1991, 1993], who related it to many issues in the philosophy of mind and language, and

Paul Churchland [1992], who also employed it as a new foundation for epistemology and philosophy of science. Yet, many proponents of symbolic AI were unconvinced. Jerry Fodor and Zenon Pylyshyn argued that connectionism had serious limitations because its representations lacked the systematicity, productivity and inferential coherence needed for human language use and reasoning. Only representations with a syntactic and semantic structure could provide these properties, they argued [Fodor & Pylyshyn, 1988]. Proponents of connectionism either denied that cognition had the properties described by Fodor and Pylyshyn or that syntactic and semantic structure were necessary to produce them, or argued that connectionist networks could approximate or instantiate symbol systems for the modeling of language use and reasoning [Smolensky, 1988; Clark, 1989].

Having previously rejected symbolic AI, Hubert Dreyfus praised connectionism for rejecting the rationalist conception of cognition in symbolic AI, and held that its basic assumptions were compatible with his own vision of intelligence [Dreyfus, 1992; Dreyfus & Dreyfus, 1988]. Yet, Dreyfus is ultimately pessimistic about its prospects for AI, because of the incredible complexity of human intelligence. The commonsense knowledge problem applies just as much to connectionism as it does to symbolic AI. In connectionist networks, the ability to deal intelligently with new situations depends on the ability to *generalize* intelligently from past experiences to new ones. This ability, Dreyfus argues, requires significant amounts of background knowledge. A neural network with such background knowledge would have to consist of millions or billions of processors, not the tens or hundreds found in most current networks. Acquisition of such knowledge would moreover require extended embodied interaction with an environment, whereas neural networks are still essentially disembodied.

'Artificial Life' shares the biological underpinning of connectionism, but rather than taking the brain's processing power as inspiration, it takes the evolutionary processes that have created the brain as its source of inspiration. Thus, whereas AI in practice involves computer simulations of human-like intelligence, Artificial Life involves computer simulations of life and life-like processes [Bedau, 2003]. One of the key differences is that the mechanisms of life are better known and easier to conceptualize than intelligence (let alone consciousness). Two further distinguishing features of ALife research is that it tends to focus on the *essential* features of living systems and to understand such systems by artificially synthesizing extremely simple forms of them (cf. Bedau [2004]). Keeley [1998] claims that one can identify a strong and a weak version of ALife, in which the weak version holds that computers and/or robots can be designed in such a way that they can be effectively used as tools in the formulation and testing of biological theories. The strong version goes further by arguing that such systems could actually be considered as being biological and alive. Boden [1996] also describes ALife as an eclectic endeavor, in which researchers are interested in very different results, ranging from the development of more efficient computational algorithms to better understanding the foundations, possibilities and

limitations of biological life. ALife can also be used to shed light on phenomena that emerge as a result of complex networks beyond biological life, such as can be found in economics and other social phenomena.

Bedau argues that one of the most important differences between symbolic AI and ALife is that the former is top-down and the latter is bottom-up. That is, symbolic AI models involve a central controller that monitors the system's global state and makes decisions that affects any aspect of the system. ALife, however, tends to take a bottom-up approach, utilizing simple "stupid" agents that interact with each other and only together determine the state and behavior of the system. In this regard, ALife shares many of the characteristic features of connectionism and Brooks' anti-representationalism (see Section 4.5) by removing the necessity for centralized processing and explicit representation. Several other aspects of ALife are interesting for philosophical research: the conceptualization/definition of life, the possibility or impossibility of creating life in a digital computer, the relationship between cognition and life (which relates to the relationship between ALife and AI) and the ethical implications of creating artificial life. Dennett [1994] has argued that ALife also ought to be regarded as a method for doing philosophy; ALife can impose requirements on thought experiments that could never be imposed by reasoning alone, thereby yielding new insights about their feasibility and possible implications.

ALife also involves the use of notions from biology to improve computing. The latter is referred to as evolutionary computing, which is a collective term for a range of different approaches that are based on principles of biological evolution. Evolutionary computation inherits many of the characteristics of natural evolution, in particular the ability to provide good, although usually not optimal, solutions on the basis of trial-and-error and some means of reinforcing the "fittest" solution. Thus, evolutionary computing is often the most viable alternative for algorithms that do not have to deliver optimal solutions (e.g. when optimal solutions are intractable) and that are applicable to a wide range of (often unforeseeable) problems (cf. Eiben & Smith [2003]).

Finally, a recent trend in AI research and cognitive science has been increased focus on dynamical systems. Dynamical systems theory focuses on how all aspects of a system can be seen as changing from one total state to another [Port and van Gelder, 1995, 15]. In other words, DST stresses this kind of holism as an alternative to the modularity, representation and (especially sequential) computation found in traditional approaches to AI. It is primarily a theory of cognitive development and the theory itself need not take an explicit stand on the possibility of its realization in a computer [van Gelder, 2000, p. 9]. However, the approach does put new items on the agenda for AI researchers by further opposing representationalism and emphasizing the importance of the brain, body and environment as a dynamical, holistic system.

4.4 Knowledge Engineering and Expert Systems

Knowledge engineering is the task of transferring human knowledge into a database (cf. Section 3.3) that can serve as a basis for enabling AI applications to perform human-like reasoning. Although there are many different kinds of knowledge engineering, we can make a rough distinction between the engineering of common sense knowledge and of expert knowledge. Research into the former was especially fueled by Hayes' seminal paper on "The naïve physics manifesto", in which Hayes argues that the engineering of everyday knowledge can be done, that it needs to be done, and he outlines a way of getting it done [Hayes, 1990]. One of the most famous examples of a common sense knowledge base is Lenat's ambitious CYC-project (cf. Lenat and Guha [1990]). The CYC project aims to eventually have a suitable representation for the full range of human expression, so that *expert* knowledge bases can be created with CYC as its basis. Despite some success (cf. www.cyc.com) the endeavor has been heavily criticized. For instance, Drew McDermott, who for a long time was an avid supporter of Hayes' program, recanted and argued that the approach commits the "logicist" fallacy of assuming that all human reasoning is necessarily deductive [McDermott, 1990].

A somewhat different approach to knowledge engineering is to focus on a limited domain of knowledge and try to understand and represent expert knowledge and reasoning. Expert systems, the first of which were developed in the middle of the 1970s, are computer systems which are intended to take over tasks from human experts in a particular specialized domain, for instance in medicine, law, mathematics and financial planning. Such expert systems typically include a knowledgebase of expert knowledge and advanced artificial intelligence to ensure that the system returns reliable and accurate answers in response to non-experts' queries. Expert systems are mainly built according to the assumptions of symbolic AI. Their designers try to provide these systems with the required knowledge by interviewing experts with the goal of making their often tacit knowledge explicit and arrive at formal rules that experts are thought to follow. The quality of expert systems is usually assessed by comparing its performance with that of a human expert.

Despite his criticism of symbolic AI, Dreyfus was relatively optimistic in his early work about the prospects of expert systems because of the formal nature of much expert reasoning. Later, Dreyfus famously reconsidered his view, concluding that humans do employ rules in early stages of learning, but that real experts replace this with an intuitive and holistic manner of problem solving [Dreyfus & Dreyfus, 1984; 1986]. In more philosophical terms, Dreyfus refuses both the epistemological and ontological assumptions behind expert systems, arguing that neither human knowledge nor physical reality has a formal structure that can be fully described in terms of rules. He thereby echoes a similar claim made by Weizenbaum, who already in 1976 attacked the tendency to reduce human problems to calculable, logical problems and emphasized the importance of intuitive human judgment even in specialized domains [Weizenbaum, 1976].

Manning [1987] argues that it is not only formalization of knowledge that poses a problem. Many problems arise because of the interaction between the expert system and the users. For instance, expert systems cannot match human experts when it comes to asking appropriate follow-up questions (e.g. if the user query does not contain enough information), to make context-sensitive distinctions between relevant and irrelevant information (cf. Section 4.2) and to separate between information that *needs* explanation and information that can *provide* explanation (e.g. in medicine, to separate between symptoms and causes of an illness). Another problem stems from the fact that expert systems must employ some kind of probability measure, since in most cases the available knowledge and user information is only sufficient to make one result more probable than the others. The question, then, becomes whether to utilize subjective measures of probability assigned by the experts or more objective measures of probability, for instance as a function of statistical data [Gillies, 2004].

These challenges give rise to a number of limitations in the range of application of (symbolic) expert systems. If expert systems cannot make decisions or form judgments at the level of an expert, they cannot be entrusted with tasks that require expertise. However, even Dreyfus admits that expert systems can often attain a certain degree of *competence*, which is a higher performance level than a human novice or advanced beginner. Expert systems therefore might indeed prove useful in applications that do not call for performance at the expert level. The decision when this is the case is related to pragmatic concerns regarding the availability of human experts and, more importantly, ethical/legal notions of risk and responsibility (see Section 4.6).

4.5 Robots and Artificial Agents

The notion of (artificial) agency is often used in computer science to refer to a computer program that is able to act on and interact with its environment. Different sets of requirements have been proposed for what it means to be an agent, which has resulted in a complex and often inconsistent set of terms for different kinds of agency, including autonomous agents, intelligent agents, adaptive agents, mobile agents etc. Sycara argues that there are four properties that characterize an artificial agent and distinguishes it from object-oriented systems or expert systems. These are 1) *situatedness*, which means that the agent receives some form of input from its environment and can perform actions that change the environment in some way; 2) *autonomy*, which means that the agent can act without human intervention and control its own actions and states; 3) *adaptivity*, meaning that it is capable of taking initiative based on its (usually pre-programmed) goals, learn from its experiences and have a flexible repertoire of possible actions; and 4) *sociability*, referring to the ability to interact in a peer-to-peer manner with other agents or humans [Sycara, 1998, p. 11]. Insofar as these requirements are too restrictive, a taxonomy could be constructed on the basis of the satisfied conditions, respectively, ‘situated agents’, ‘autonomous agents’, ‘adaptive agents’, and ‘social agents’. Furthermore, it is important

to note that ‘environment’ should be interpreted in such a way that it includes non-physical environments such as cyberspace; arguably the most successful artificial agents are those who perform one or more of the criteria above in cyberspace – usually referred to as bots.

The notion of autonomy is also used in different senses, ranging from simple and highly specialized agents such as advanced factory robots, to complex, multi-purpose, multi-sensorial agents with the capability to learn and display behavior that would be regarded as intelligent if carried out by a human being. At a minimum, ‘autonomous’ carries *some* of its philosophical meaning in the sense that an autonomous agent should be able to make informed decisions (based on its knowledgebase, rules and sensory input) and act accordingly. However, Haselager [2007] has argued that the notion of autonomy in robotics and philosophy is radically different, referring respectively to independent performance of tasks and capacity to choose goals for oneself. He suggests that this gap can be bridged by interpreting ‘autonomy’ in terms of what it is that makes one’s goals genuinely one’s own, further emphasizing the importance of embodiment to genuine autonomy. Insofar as artificial agents need to act in a complex and dynamic world, many of the epistemological questions discussed through the history of philosophy become relevant. For instance, they should ideally be able to revise their “beliefs” in light of their experiences, which requires functions for preserving consistency between beliefs, differentiating between beliefs that require justification and those that do not and so forth. Importantly, such revision will differ depending on what kind of substantive theory of truth that lies behind. For instance, an intelligent agent will operate differently depending on whether it will revise its beliefs according to a coherentist or correspondence theory of truth. Thus, there is a close connection between epistemological problems in philosophy and robotics (cf. Lacey and Lee [2003]).

Colloquially, people tend to use the word ‘robot’ for artificial agents that have a human or animal appearance and these robots tend to produce more natural man-machine interaction. At least, this is the guiding principle behind the MIT Cog project. Cog is a humanoid robot designed to gain experience from natural interactions with humans. The guiding principles behind this project are that human-level intelligence requires social interactions akin to those of a human infant and that a humanoid robot is more likely to elicit natural interactions [Brooks, 2002]. As such, Cog and other advanced intelligent agents can be seen as a means of empirically testing the more abstract theories in philosophy of artificial intelligence. In particular, Cog illustrates a recurring theme in philosophy of AI. Many philosophers have claimed that “the nonformalizable form of information processing … is possible only for embodied beings” [Dreyfus, 1992, p.237] and that robotics stand a better chance of producing human-like intelligence. This claim, originally made by Turing more than 50 years ago, has been taken to heart by Rodney Brooks, who claims that the visionary ideas of Turing can now be taken seriously; AI should focus on robotic architectures that are inspired by biology and interact with the actual world rather than simply reason according to a set of formal rules and knowledge bases that are far removed from the

complexity that human-like intelligence must be able to handle [cf. Brooks, 2002].

4.6 AI and Ethics

If we regard the ultimate goal of AI to be machines capable of doing things that have traditionally required human intelligence, many ethical questions become obvious. For instance, can computers be trusted with tasks that involve considerable risk to others, and what are the social and existential implications of substituting man for machine to an even higher degree than we already have? One major problem with the development of autonomous systems is that it tends to erode the notion of 'responsibility', arguably the most important concept in both ethics and law. In general, it can be difficult to assign responsibility if computer malfunction results in loss of lives (see also Section 6.1). Are the designers, the users, or perhaps even the artificial agents themselves responsible? Indeed, Sullins [2006] argues that if an artificial agent were to *understand* responsibility – that is, if the only way we can make sense of its behavior is to ascribe to it an understanding of responsibility – then it should be treated as having both rights and responsibilities regardless of whether it is a person or not. Floridi and Sanders have argued that although artificial agents cannot be blamed or praised for their actions, they ought – when seen at a given level of abstraction – to be regarded as moral agents in the sense of being the sources of good or evil [Floridi and Sanders, 2004]. They further claim that regarding artificial agents as moral agents does not reduce the responsibility of the designers. On the contrary, seeing them as sources of immorality should prompt us to pay extra attention to the kinds of agency these entities have. Johnson, however, sees a danger in assigning moral agency to artificial agents, "because it disconnects computer behavior from human behavior, the human behavior that creates and deploys the computer systems" [Johnson, 2006, p. 204]. That is, contrary to Floridi and Sanders, Johnson argues that the design of artificial agents is more likely to be subject to moral scrutiny if we focus on computer systems as human-made rather than as independent moral agents. A clear distinction between humans and computers also underlies Moor's conclusion that computers should not be allowed to make decisions about our basic goals, values, and priorities between them [Moor, 1979].

As these viewpoints on artificial moral agency show, the source of many ethical problems in AI stems from the fact that these systems tends to be opaque. That is, they make *choices* and *decisions* according to criteria of which the users generally have little or no understanding. These operations can even be opaque to the designers themselves, especially when build upon a connectionist or evolutionary architecture without formal rules and representations (see also Section 3.2 on validation/verification). In Brey's terminology [2000], the most important task facing computer scientists and ethicists is to *disclose* such problems and hidden values beforehand, rather than dealing with them afterwards. With the complexity required for a machine to act intelligently, it could even be argued that it is impossible to safeguard against their malfunction

and that the creation of war robots (cf. Asaro [forthcoming]) and other AI systems capable of massive destruction is inherently unethical. Indeed, this line of reasoning famously led David Parnas to take a public stand against the so-called Star Wars program during the cold war, arguing that it would be impossible to create artificial intelligence that can reliably be trusted to prevent nuclear attacks (cf. Parnas [1985]).

Some more futuristic ethical problems have also been raised, for instance surrounding the discussion whether artificial agents can acquire moral status and rights comparable to humans, or at least simpler life forms. These kinds of questions have long been the subject of literature and movies. Many of the most fundamental problems are illustrated by Asimov's famous laws of robotics, which state that 1) a robot may not injure a human being, or, through inaction, allow a human being to come to harm; 2) a robot must obey orders given it by human beings, except where such orders would conflict with the First Law; 3) a robot must protect its own existence as long as such protection does not conflict with the First or Second Law [Asimov, 1968]. The ethical problems that arise in Asimov's works of fiction typically arise as a result of irresolvable conflicts between these laws, illustrating the difficulty with which ethical guidelines can be formalized. The laws also presuppose that robots have mere instrumental value and consequently should be regarded as means and not as ends in themselves. This raises the question of what it would do to our sense of humanity if machines were to become better at reasoning than humans, rationality having traditionally been seen as the very essence of humanity (see e.g. Mazlish [1993]). Furthermore, Asimov's laws illustrate the two of the biggest problems in creating the kinds of robots that would need such laws. First, such a robot must be able to make nuanced and reliable distinctions between moral patients, non-moral patients – and its own being. Second, it must be able to understand the consequences of its actions – and *inactions.* Allen, Smit and Wallach suggest that the latter can be done either in a top-down approach, which involves turning moral theories into algorithms, or bottom-up, which involves attempts to train artificial agents in such a way that their behavior emulates morally praiseworthy human behavior (Allen, Smit and Wallach [2005]; see also Clarke [1994]). However, regardless of how successful we are in trying to create artificial morality, the mere attempt can be advantageous for many of the same reasons that AI in general could lead to new insights without necessarily leading to success. As Knuth puts it, in speaking of the impact of computers on mathematics, the mere *attempt* to "formalize things as algorithms leads to a much deeper understanding than if we simply try to understand things in the traditional way" [Knuth, 1974a, p. 327; cf. Gips, 1994]

Finally, if computers become reliable to such a degree that we willingly leave our deliberations and decisions to the computer, does this entail that *our* autonomy is reduced? Perhaps the first to raise these kinds of issues was Joseph Weizenbaum [1976], who himself had created the famous artificial therapist ELIZA. Weizenbaum became increasingly worried about the

effects of ELIZA, especially due to the fact that people confided in it despite knowing that it was a computer program and that psychologists considered developing it further and putting it into real practice. Weizenbaum argued that the question is not *what* intelligent machines can do, but *whether* we should allow them to do it. In other words, the guiding principle behind AI research should not be determined by a form of technological determinism in which we do something simply because we can. Ethics is not a question that should be *raised* by AI, but it should be the very foundation of AI; The justification for building a system in the first place should be an ethical question.

5. Philosophy of the Internet and New Media

This section discusses issues and approaches in the newly emerged field of philosophy of the Internet and new media, sometimes called cyberphilosophy, which has emerged together with the multidisciplinary field of new media studies. The first section will give a broad outline of new media, Internet in particular, and discuss theories on how society has increasingly become an information society. In the two subsequent sections, we consider epistemological and ontological issues relating to the Internet and other new media. Section 5.4 considers new media as a platform for communication and the establishment of virtual communities, followed by a related section on the Internet as a political venue. The chapter concludes with a section on how our identity is affected by the disappearing barriers between body and technology and between real and virtual selves. Ethical issues will be considered occasionally throughout, but will also be discussed separately in Section 6.

5.1 Theories of New Media and the Information Society

The emergence of multimedia computers in the 1980s and the Internet as a mass medium in the early 1990s created a new role for computer technology. This development moved computers beyond scientific and administrative and organizational applications, and made them into a social medium and mass medium – a general-purpose tool and environment for games, creative expression, art, film and photography etc. Furthermore, networked computers, as facilitated by the Internet, allowed individuals to communicate and perform joint activities over computers.

'New media' generally refers to the recent forms of media that rely on digital computers, including both the development of unique forms of digital media, such as virtual worlds, and the transformation of traditional media, such as movies streamed on demand on the Internet [Flew, 2002]. The development of new media is also closely related to the development of increasingly mobile, ubiquitous and interconnected devices, which enable access to new media at any time and any place. Another important feature of new media is its facilitation of user contributions; Users can for instance generate and share original content, find and retrieve content on demand,

or publicly rate, comment and recommend content. As such, new media often focus on “the relationship between an individual, who is both a sender and receiver, and the mediated environment with which he or she interacts” [Steuer, 1995, p. 7]. In other words, new media is often a form of many-to-many communication. This cannot be achieved by simply *transferring* (or *pushing*) information simultaneously from many to many. Instead, a venue must be created in which many can leave information and many can retrieve (or *pull)* information; anything else would amount to chaos and information overload. Thus, traditional forms of media are sometimes described as *channels* of information, but a more apt analogy for new media is a *place for* information – which is reflected in terms like cyberspace, infosphere, virtual worlds and virtual environments. This form of interactivity entails that users are not left with a choice between 'on' or 'off', but also what, when and how. Thus, new media has increasingly relied upon a community of users and often place emphasis on *sharing* and *collaboration* – what has also been referred to as a bazaar rather than a cathedral model [Raymond, 2000]. This societal model and the increasing importance of information technology in our lives, especially evidenced by the Internet, have reinforced the characterization of modern society as an ‘information society’.

There is now considerable agreement among social theorists, economists and historians that contemporary society can be characterized as an information society, which is in important respects different from the industrial society that preceded it until the 1970s [Webster, 1995]. The information society is a society in which the production and use of information has a dominant role in economic and social processes. In the economy of the information society, it is the information industry rather than the industry of goods that is the driving force. The contemporary economy is dominated by companies in the areas of communication, media, IT, advertising, and information services. Information and information technology have also become more important in traditional industry and services. Socially and culturally, the transition to an information society has also introduced major changes in work, leisure, social organization and lifestyles.

According to Manuel Castells, who has presented the most comprehensive theory of the information society to date, the information society is the result of a transformation of the capitalist economy through information technology, which has made capitalism more flexible, more global and more self-sustained. In this new model, the basic unit of economic organization is no longer the factory but the network, made up of subjects and organizations and continually modified in adaptation to its (market) environment. Castells argues that contemporary society is characterized by a bipolar opposition between the Net (the abstract universalism of global networks) and the Self (the strategies by which people try to affirm their identities), an opposition which is the source of new forms of social struggle [Castells, 1996].

The transition to an information society is also theorized by Van Dijk [2006], who argues that the information revolution in the 1970s was preceded by a crisis of control in organizations, which were held back by uncontrolled bureaucracy, limitations in transportation systems, and

inadequacies of mass communication in an individualizing and diversifying society. He argues that new media technologies enabled a revolution in information and communication that helped solve these problems and enbabled the transition from a Fordist to a postfordist mode of production in which organizations become more streamlined and flexible and better able to operate on a global scale.

These characterizations of the transition of an industrial to a postindustrial information society is accepted by postmodern thinkers like David Harvey, Frederick Jameson, Jean Baudrillard, Mark Poster, and Jean- François Lyotard, who add that these technological and economic changes are accompanied with distinct social, cultural and epistemological changes which designate a shift from a modern to a postmodern culture and society. The new information-based capitalism has engendered new patterns of consumption, lifestyles, modes of social organization and association and patterns of cognition and experience. In general, postmodern authors characterize the information society as a society in which modern life has become saturated by information, signals and media, in which there is a decline of epistemic and political authorities, and which is characterized by consumerism, commodification, simulation, a blurring of the distinction between representation and reality, and the fragmentation of experience and personal identity.

Harvey [1989] has argued that the new economic system has led to a new dynamism in which work and consumption are sped up and made more competitive, and a new, postmodern culture which rejects the faith in reason and objective reality and accepts heterogeneity and commodification. Paul Virilio [1994] holds a similar but more somber view. He holds that the marriage of capitalism and new (media) technologies have created a culture of speed, which ultimately leads to a feeling of incarceration and confinement in the world. Jean Baudrillard [1995] theorizes a shift from an economy of goods to an economy of signs and spaces and characterizes the new era as an era of simulation, rather than information. He holds that the new social order is determined by models, signs and codes, and leads to a disappearance of the distinction between representation and reality, past and future, catching people in a disorienting postmodern hyperspace. Baudrillard claims, along with Poster [1990] and Lyotard [1984] that contemporary life is ruled by a new ‘mode of information,’ in which life is quintessentially about symbolisation, about exchanging and receiving messages about ourselves and others. Lyotard claims, in addition, that postmodern society is characterized by the commodification of knowledge, which has become decentralized and made accessible to laypersons.

The emergence of the Internet as a mass medium in the 1990s has further strengthened the arguments of theorists of the Information society. However, it has also left some of the older theories dated, due to the explosive development of the Internet. One notable exception is Floridi, who argues that we are probably the last generation to experience a clear difference between offline and online. The blurring and eventual dissolution of the two, Floridi argues, stem from

three fundamental trends. First, there has been a steady increase in the *kinds* of information that can be represented digitally, already encompassing all other forms of media. The same holds for the *amount* of information produced. Second, technologies like Radio Frequency Identifiers (RFID) will increasingly allow us to be continuously online and to communicate with non-living entities like cookware and clothing (and for them to communicate with each other). Third, the information society is becoming a collection of "connected information organisms" or 'Inforgs' and we are about to become the "only biological species capable of creating a synthetic environment to which it then must adapt" [Floridi, 2006]. Floridi's vision emphasizes the emergence of new and pervasive forms of connectedness, between humans and humans, humans and machines, and machines and machines, all mediated by the flow of information. How the new form of sociality this entails for humans differs from – and whether it is less valuable than – traditional forms gives rise to many of the issues discussed in the following sections.

5.2 Internet Epistemology

The Internet is a global tool for the production, storage, dissemination, and consumption of information and knowledge on which a large percentage of the world population relies. Given the dominant role of the Internet as both a source of information and a means for the production of information, an investigation of its epistemic properties is warranted. Internet epistemology is an emerging area of applied epistemology that evaluates epistemic properties of Internet technology and practices of information production, management and utilization on the Internet. In addition it could propose improved epistemic practices and technologies. The term "Internet epistemology" was first introduced by Paul Thagard [2001] to refer to the epistemology of scientific information practices on the Internet, but is now gaining a wider usage to include everyday information practices as well. Issues in Internet epistemology include the epistemic quality of Internet information, the normative implications of the Internet for information production and consumption and the epistemology of Internet-related information practices, including information utilization, management and production.

The quality of information on the Internet, and the epistemic value of the Internet as a source of information, has been questioned by Alvin Goldman [1999, 161-189]. Goldman argues that while the Internet strongly increases the amount of information available to us, it need not be true that we know more as a result. The expansion of knowledge depends on both the truthfulness and the usefulness of the content produced and the ability of users to distinguish true and relevant from false and irrelevant content. The quality of the Internet as a source of information thus depends on both its support for intelligent information retrieval that gives users access to information relevant to their interests, and the reliability of the information that is thus made accessible. The first issue will be referred to as the *problem of relevance*, while the second will be called the *problem of reliability*.

It is generally agreed that the *problem of relevance* has not yet been solved adequately. The Internet is often criticized for offering users vast amounts of information without sufficient means for identifying and retrieving the information that is most relevant to their interests. This plethora of unorganized information has been argued to contribute to *information overload*, or information glut, which is a condition in which recipients are being provided with more information than they can meaningfully process [Shenk, 1998; Himma, 2007b]. Information overload leads to information fatigue and to recipients becoming paralyzed in their decision-making capabilities or to them remaining uninformed about topics. Proper information management, both by information providers and recipients, can provide a solution to information overload. Tagging and categorization of web documents and sites, search engines, hierarchical directories, filtering techniques, hyperlinking, personalized information retrieval profiles, and the development of semantic web technologies can further facilitate information retrieval. Many of these techniques depend on automated procedures. Dreyfus [2001, p. 8-26] argues that these will ultimately fail because computers are insufficiently capable of discerning relevant from irrelevant information (see also Section 4.2). Levy [2008] argues that good information management is not enough to avoid information overload, and that the creation of space and time for thinking, reflection and extensive reading is necessary as well.

The *problem of reliability* also looms large for the Internet. Issues include the fact that anyone can place whatever information on the Internet that they please, that the source of information on the Internet is often unclear, and that it is often unclear whether information on the Internet is still current. Furthermore, websites usually lack information on criteria used in the selection of the information provided or referenced on them. The problem of reliability of Internet information has been addressed by Anton Vedder [2008, Vedder & Wachbroit, 2003]. Vedder argues that generally, persons evaluate the reliability of information presented to them by means of two types of criteria: *content criteria* and *pedigree criteria*. Content criteria are criteria of reliability inferred from the information itself. They include criteria of consistency, coherence, accuracy, accordance with observations, and interpretability, accessibility, and applicability relative to the user's capacities and interests. Pedigree criteria are epistemic criteria to assess the authority, trustworthiness and credibility of the persons or organizations *behind* the information.

Vedder argues that people often cannot determine the reliability of information on content criteria alone. Reliability is often evaluated in large part through an evaluation of the epistemic authority and credibility of the source of information, using pedigree criteria. Pedigree information is often not available in Internet information, so that according to Vedder recipients are dependent on content criteria alone. Two such criteria have become dominant: accessibility and usability. Many users choose to rely on Internet information solely based on it being easily available and applicable for their purposes. Vedder argues that this undesirable state-of-affairs can be remedied through two strategies: developing critical attitudes in recipients and making pedigree

criteria visible on the Internet by creating institutionally licensed seals and procedures for verifying the credibility of websites. Fallis [2004] offers a somewhat similar argument based on the epistemology of testimony.

In addition to the quality of Internet information, philosophers have also analyzed and evaluated the wider implications of the Internet for information production and consumption. One prominent author to do so is Luciano Floridi [1999, 81-87], who outlines eleven implications of the Internet for organized knowledge, including the decentralization and possible fragmentation of knowledge, a potential blurring of the distinction between information and knowledge, the emergence of the computerized scholar, and the loss of information on paper. He argues that we are "giving the body of organized knowledge a new electronic life" and argues that we must do so carefully and wisely.

Hubert Dreyfus [1999, 10-11] argues that we are moving from a library culture, built on classification, careful selection and permanent collections, to a hyperlinked culture, involving diversification, access to everything, and dynamic collections. The old library culture presupposes a modern subject with a fixed identity seeking a more complete and reliable model of the world, whereas the new hyperlinked culture assumes a postmodern subject not interested in collecting and selecting but in connecting to whatever information is out there. Dreyfus claims that the old information culture is superior to the new one; poststructuralists and postmodernists would argue the opposite. Similar conservative positions are taken by Albert Borgmann [1999], who argues that digital information creates an alternate reality, which is ambiguous, chaotic and fragile, and could threaten traditional modes of information, and Phil Mullins [1996], who worries that in a culture which relies on electronic documents and hypertext, books become fluid and decentered, and canons dissolve. Mullins thereby echoes an earlier claim made by Walter Benjamin to the effect that the ease of reproduction offered by modern technology leads to detachment from, and therefore shattering of, tradition [Benjamin, 2006].

Authors who study the organization of information on the Internet agree that the dominant mode of organization is hypertextual. Hypertext is text which contains dynamic links, called hyperlinks, to other texts. When using a text, users can retrieve related texts, by clicking the relevant hyperlink. Dreyfus [1999, 8-26] argues that hyperlinks link pieces of information to each other based on some subjective perceived relationship, removing hierarchy and authority from the organization of information, and with it, meaningfulness. Hyperlinks, he argues, are not designed to support the retrieval of meaningful and useful information, but rather to support lifestyles oriented at surprise and wonder. Floridi [1999, pp.116-131) presents an opposite point of view, arguing that hypertext significantly contributes to the meaningfulness of information on the Internet by providing semantic structure between its separate texts. Hypertext, he argues, allows for multi-linear narratives and a more open and flexible space of knowledge which he does not characterize as postmodern, but rather as marking a return of the Renaissance mind.

Information and knowledge on the internet can be produced offline, by individuals and collectives, and then put online, or it can be produced online, both individually or, epistemologically most interesting, through collaboration. The Internet has become a medium for collaborative knowledge creation, enabling the utilization of both synchronous (i.e., real-time) media such as real-time groupware, instant messaging, and multi-user editors, and asynchronous media such as email, version control systems, open source software development tools, wiki's and blogs. What are the epistemological properties of both these media and the associated collaborative practices, and how can we compare the epistemic quality of alternative collaborative practices? These questions fall within the scope of social epistemology [Goldman, 1999], which is the study of the social dimensions of knowledge or information.

The most innovative development in knowledge production on the Internet is the emergence of nonhierarchical forms of mass collaboration, including the creation of wiki's, open source software design, and collaborative blogging [Tapscott and Williams, 2008; Sunstein, 2006]. These practices are community-based, voluntary, egalitarian, and based on self-organization rather than top-down control. The relative success of these practices seems to show that such systems of knowledge creation can be as successful as more traditional systems. Fallis [2008] argues that Wikipedia, the online collaborative encyclopedia, is reasonably reliable and has additional epistemic virtues that recommend it as a source of knowledge. Goldman [2008] compares political blogging to traditional media news, and argues that it can be a reliable and informative source of information provided that the bloggers are sufficiently motivated to bring out a political message. Thagard [2001] considers how scientific knowledge production is being reshaped by the Internet. He considers various Internet technologies that have become part of scientific practice, like preprint archives, newsgroups and online databases, and evaluates them for their epistemic quality using a set of epistemic criteria proposed by Goldman [1986].

Turning to knowledge utilization, one development that merits attention is distance learning, or distance education. Hubert Dreyfus [1999] has argued that education centrally involves the transmission of skills and the fostering of commitment by educators in students to develop strong identities. According to Dreyfus such aspects of education cannot adequately be transferred in distance education since they require bodily presence and localized interactions between students and teachers. Prosser and Ward [2000] add that the transfer of "practical wisdom" in education requires communities with interpersonal connectivity among its members, something virtual communities in distance education cannot provide because of their relative anonymity, the lack of mutual commitments, and the risk of an overload of trivial information. Nissenbaum and Walker [1998] provide a more nuanced view, arguing that the implications of information technology for learning depend on the actions and attitudes of instructors and policy makers.

5.3 The Ontology of Cyberspace and Virtual Reality

The software constructs with which computer users interact, such as files, folders, and web pages, exist virtually rather than physically. Although they are realized and sustained by means of physical systems, they do not exist solely or primarily as physical entities. A web page, for example, does not appear to have physical properties, such as mass, weight, and coordinates in physical space. It may, however, have a location in virtual space, given by its internet address, and it has perceptual, formal and logical properties by which it can be identified. The existence of virtual objects, and correlated virtual spaces, actions and events raises questions regarding their ontological status: what is their mode of existence, and what is their place in philosophical ontology?

Let us call nonphysical software-generated objects and spaces with which users interact *virtual entities*. Virtual entities are represented as part of the *user interface* of computer programs. They are part of an ontology defined by the program that specifies classes of objects with which the user interacts. They manifest themselves to the user through symbolic or graphical representations, and they interactively respond to actions of the user. Contemporary user interfaces are in most cases *graphical*, representing virtual objects as ordinary, manipulable physical objects. *Virtual reality*(VR) is special kind of graphical user interface which presents a computer-generated immersive, three-dimensional, interactive environment that is accessed and manipulated using, for instance, stereo headphones, head-mounted stereo television goggles, and datagloves. VR technology, which can be single- and multi-user ("networked VR") allows for a representation of virtual entities with a great degree of realism.

The Internet and other computer networks define collective, multi-user virtual entities in a collective user environment that is often called *cyberspace* [Benedikt, 1991]. Although cyberspace is accessed using graphical user interfaces, most virtual entities it contains are informational objects like web pages, text and image files, and video documents. Located conceptually in between cyberspace and VR one finds *virtual environments* or *virtual worlds*. Although the term virtual environment is sometimes used synonymously with virtual reality, it more often is used to denote any interactive computer simulation of an environment, whether represented textually or graphically, and whether immersive or nonimmersive, which can be navigated by users. (For a discussion of these and related conceptual distinctions, see Brey [2008] and Brey [1998]).

Brey [1998] argues that the analogy between virtual and physical spaces goes deep because both are topological spaces as defined by mathematical topology. Topology is a branch of mathematics that defines topological spaces as mathematical structures that define abstract relations of closeness and connectedness between objects in terms of relationships between sets rather than geometrical properties. A directory can be said to contain a file if the right topological

relation exists between them, and webspace can be defined as the topological space generated by the system of hyperlinks between pages on the web.

This characterization of virtual space does not yet answer our question regarding the ontological status of virtual entities. A common conception is that “virtual” contrasts with “real” and that therefore virtual entities are not real in an ontologically robust sense. They are hence constructions of the mind, or mere representations. Borgmann [1999] argues that virtual reality is therefore always only a make-believe reality, and can as such be used for entertainment or training, but it would be a big mistake to call anything in virtual reality real, and to start treating it as such [Borgmann, 1999]. Philip Zhai [1998] takes a radically opposing point of view, arguing that something is real when it is meaningful to us, and that consequently there is no principled ontological distinction between virtual and physical reality.

Steering in between idealist and realist conceptions of virtual entities, Brey [2003] has argued that virtual is not the opposite of real, and that some but not all virtual entities are virtual and real at the same time. Brey argues that a distinction can be made between two types of virtual entities: simulations and ontological reproductions. *Simulations* are virtual versions of real-world entities that have a perceptual or functional similarity to them, but do not have the pragmatic worth or effects of the corresponding real-world equivalent. *Ontological reproductions* are computer simulations of real-world entities that have (nearly) the same value or pragmatic effects as their real world counterparts. He argues that two classes of physical objects and processes can be ontologically reproduced on computers. A first class consists of physical entities that are defined in terms of visual, auditory or computational properties that can be fully realized on multimedia computers, such as images, movies, musical pieces, stereo systems and calculators.

A second class consists of what John Searle [1995] has called *institutional entities*, which are entities that are defined by a status or function that has been assigned to them within a social institution or practice. Examples of institutional entities are activities like buying, voting, owning, chatting, playing chess, trespassing and joining a club, and requisite objects like contracts, money, chat rooms, letters and chess pieces. Most institutional entities are not dependent on a physical medium, because they are only dependent on the collective assignment of a status or function. For this reason, many of our institutions and correlated practices and objects, whether social, cultural, religious or economic, can exist in virtual or electronic form. Institutional virtual entities are the focus of David Koepsell [2000], who critiques the existing legal ontology of entities in cyberspace. Because virtual entities may have different ontological statuses, ranging from fictional (virtual oranges and some virtual cheques) to real (other virtual cheques and virtual chess games), and because some entities can change their ontological status through the collective assignment of functions, users of cyberspace and virtual may encounter ontological confusion.

Although Brey blurs the distinction between reality and virtuality, he does maintain a distinction between reality and fiction. Some authors have however argued that the distinction between simulation and reality and thus truth and fiction is being erased with the emergence of computer-generated realities. Jean Baudrillard [1995] has claimed that information technology, media, and cybernetics have ushered in an era of simulation, in which models, signs and codes mediate access to reality and define it to the extent that people cannot meaningfully distinguish between simulations and reality anymore. Albert Borgmann [1999] has argued that virtual reality and cyberspace have incorrectly led many people to confuse them for alternative realities that have the same actuality as the real world, whereas he holds that VR and cyberspace are merely forms of information that should be treated as such.

Next to these ontological and epistemological questions regarding the distinction between the virtual and the real, there is the moral question of the goodness of virtuality [Brey, 2008]. First of all, are virtual things better or worse, more or less valuable, than their physical counterparts? Some authors have argued that they are in some ways better: they tend to be more beautiful, shiny and clean, and more controllable, predictable, and timeless. They attain, as Michael Heim [1993] has argued, a supervivid hyper-reality, like the ideal forms of Platonism, more perfect and permanent than the everyday physical world. Critics have argued that these shiny objects are mere surrogates that lack authenticity [Borgmann, 1999] and that presence in VR and cyberspace gives a disembodied and therefore false experience of reality and present one with impoverished experiences [Dreyfus, 2001]. More optimistically, Mooradian [2006] claims that virtual environments and entities are good at creating hedonistic value as well as certain types of perfectionist value, notably intellectual and aesthetic value, though not value located in excellent physical activities.

5.4 Computer-Mediated Communication and Virtual Communities

The Internet has become a medium for communication and social interaction, and an increasing part of social life is now taking place online. The study of online social life is being undertaken in a number of new and overlapping interdisciplinary fields that include new media studies, Internet studies, cyberculture studies, cyberpsychology and computer-mediated communication. In philosophy, online social life has been studied within philosophy of computing and computer ethics, but the relevant parent discipline is social philosophy, which is the philosophical study of behavior, social structure and social institutions.

Let us first consider philosophical issues in computer-mediated communication. Computer-mediated communication (CMC) is interactive communication across two or more networked computers. It includes both synchronous and asynchronous communication, and such formats as e-mail, instant messaging, chatrooms, bulletin boards, listservs, MUDs, blogs, video conferencing, shared virtual reality and software-supported social networking. The field of CMC

studies these different forms of communication as well as their social effects. Philosophical studies of CMC are quite diverse, including issues like online identity, virtual public spheres, Internet pornography, and power relations in cyberspace, and thus seem to consider many issues in the philosophical study of social life online if not the Internet at large [Ess, 1996; Ess, 2004].

A key philosophical issue for CMC is how different types of CMC formats can be normatively evaluated. Is CMC epistemically and socially inferior to offline communication, including face-to-face communication, or is it in some ways superior? Shank and Cunningham [1996] argue that many forms of CMC involve multiloguing, which is unscripted, simultaneous, non-hierarchical conversation involving multiple participants. They argue that multiloguing supports diversity in perspectives, integration of knowledge, equality in participation, and access to archived dialogue, and therefore presents a superior mode of communication. Dreyfus [2001] takes a more negative view, arguing that important qualities are lost in CMC, including the movements and expressions of the body, a sense of context, and genuine commitment and risk-taking. Another major philosophical topic in CMC is communication across different cultures and worldviews [Ess, 2002; Ess and Sudweeks, 2005]. Ess [2002] has argued that cross-cultural studies of CMC can help resolve long-standing questions about the nature of culture, knowledge, politics, and the self.

Another important development is the formation of online social relationships. The Internet has become a major site for social networking, using media like e-mail, instant messaging, and social networking sites. It is being used to build up networks of acquaintances, and to forge and maintain friendships and even love relationships. A question for philosophy is how this development and the resulting new types of relationships should be evaluated. One central issue is whether online social relationships can include mutual trust [Weckert, 2005; Nissenbaum, 2001]. Pettit [2004] argues that genuine trust (as opposed to mere reliance) is not possible in exclusively online relationships because the Internet does not sufficiently support the justification of beliefs in loyalty and the communication of trust. De Laat [2005] presents an opposing view, arguing that enough social and institutional cues can be used online to develop trust.

Cocking [2008] argues that fully computer-mediated personal relationships cannot be as rich and genuine as offline relationships because people have too much control over our self-presentation online. Briggle [2008a] takes issue with this position, arguing that the Internet is well-suited for fostering close friendships based on mutual self-exploration because it creates distance and supports deliberate behavior. Ben-Ze'ev [2004] argues that the Internet enhances love relationships because it allows for meaningful online relationships in which people can express themselves very directly and in which they can live out interactive fantasies. Briggle [2008b] presents a general framework for the interpretation and evaluation of different types of online love relationships. Brey [1998], finally, criticizes the increasing substitution of social interaction by

interaction with software agents and the proliferation of social relationships with virtual characters and pets.

People do not only form individual social relationships online, they also form online communities, or v*irtual communities*, that have their existence in cyberspace. Virtual communities can be closed or open and may be intended to maintain existing relationships, or explore various kins of shared interests. It has been questioned whether virtual communities constitute genuine communities and whether they are inferior to traditional, local communities, whether they effectuate social integration or fragmentation, and whether they cause a harmful erosion of traditional, local communities [Feenberg and Barney, 2004]. Many authors have defended virtual communities, arguing that they can embody all the qualities of traditional communities, including mutual trust, care and a sense of belonging [Rheingold, 1993]. Virtual communities have been assessed mostly positively by postmodern philosophers like Lyotard and Bolter because of the non-Cartesian, decentered, fragmented, and hypertextual nature of the identities portrayed by their users (cf. Section 5.6).

Others have argued that virtual communities are inferior to traditional ones. Borgmann makes a distinction between instrumental, commodified and final communities and argues that virtual communities can at best be instrumental or commodified, because they do not contain “the fullness of reality, the bodily presence of persons and the commanding presence of things” found in final communities [Borgmann, 2004, p. 63]. In a similar fashion Barney [2004] sees virtual communities as inferior due to their lack of physical practices, and Dreyfus is critical of what he describes as the nihilist, irresponsible and often uninformed nature of virtual communities [Dreyfus, 2004]. Winner, finally, has criticized the fact that any kind of online network is called a community, since this broad definition ignores the importance of “obligations, responsibilities, constraints, and mounds of sheer work that real communities involve” [Winner, 1997, p. 17]. Interestingly, both Rheingold and Bolter have recently also adopted more conservative positions on virtual communities.

Does the proliferation of virtual communities and online social networks support social integration or does it lead to social fragmentation? Many years before the Internet, Marshall McLuhan [1962] already claimed that electronic mass media were bringing about a *global village* in which people are globally interconnected by electronic communications. It has subsequently been claimed that the Internet, more than other electronic media, has instantiated a global village. This view has met with serious criticism. The notion of a global village suggests civic engagement and a unified public sphere. Instead, Robert Putnam [2001] has argued, the ubiquitous creation of interest-based online communities has brought about a *cyberbalkanization* of online social life. He defines cyberbalkanization as a process in which cyberspace is divided into narrowly focused groups of individuals with shared views and experiences, that cut themselves off from alternative views and critique. A similar view is presented by Sunstein [2001], who emphasizes that this

process is not only caused by the creation of interest-based virtual communities, but also by the increasing ability to individually filter information on the Internet in accordance with one's previously formed beliefs.

5.5 The Internet and Politics

Political philosophy studies what kinds of political institutions we should have. It analyzes, criticizes and defends major political ideologies like liberalism, socialism, and conservatism, and tries to give content to central concepts in political theory like power, liberty, democracy and the state. The Internet is becoming an object of study for political philosophy for two reasons. First, the Internet is becoming an important means by which politics is pursued. It is being used for political dialogue between politicians and citizens, for political organization and activism, for electronic voting, for political reporting, and even for terrorist attacks. The use of the Internet for political activity has been termed *cyberpolitics*. A political philosophy of the modern state that does not take the existence of cyberpolitics into account runs the risk of using an outdated conception of the political process. In addition, cyberpolitics itself is a worthy object of study, since legitimate questions can be raised concerning the way cyberpolitics ought to be conducted.

The second reason that the Internet ought to be studied by political philosophy is because of the emergence of virtual communities and social networks in cyberspace (Section 5.4). These social structures have emerged in a medium, cyberspace, that is not subjected to the political authority of any nation of conglomerate of nations. Cyberspace has therefore been called a stateless society. Few political institutions exist within it with the authority to govern and regulate social activity. Nevertheless, cyberspace has a politics; it has processes by which individuals and groups negotiate conflicts of interests and attempt to exercise power and authority. A question for political philosophy is what the *politics of cyberspace* ought to be, that is, what political institutions and regulations ought to be in place in cyberspace. Although the politics of cyberspace and cyberpolitics are conceptually distinct, their relation should also be considered. The way in which cyberspace is organized politically may have serious consequences for the extent to which it can be used as a means for politics in the "real" world by different groups. Conversely, agents that use the Internet for "real-world" politics may adopt a presence in cyberspace and establish interactions with other agents in it, thereby becoming part of the social fabric of cyberspace and hence of its politics.

The politics of cyberspace have been an issue long before the emergence of cyberpolitics. Early pioneers of the Internet considered it a free realm, a new "electronic frontier" not subjected to laws, and they generally wanted to keep it that way. They emphasized the protection of individual and civil rights in cyberspace, such as the right to privacy, free speech, freedom of association and free enterprise (see also 6.2). As Langdon Winner [1997] has argued, the dominant political ideology of Internet users and authors in the 1980s and 1990s was

cyberlibertarianism, which he defines as an ideology that embraces radical individualism, including strong conceptions of individual rights along with an embrace of free-market capitalism, a rejection of regulative structures and a great optimism about technology as a means for liberation. Winner criticizes cyberlibertarianism for its neglect of issues of justice and equality, and its conceptions of citizenship and community. In its place, he proposes *cybercommunitarianism*, in which political structures are put in place that support communities rather than individuals.

Another issue concerns the democratic nature of cyberspace. Can and should the Internet be a democratic medium? Some have argued that the Internet is inherently a democratic technology, as it is designed as consisting of a network of equal nodes with equal opportunities for sending and receiving information. This design obliterates hierarchies, it has been claimed, and supports direct, participatory democratic processes. Deborah Johnson [1997] cautions that although the Internet can indeed empower people, the filtering of information by authorities and the insulation from diverse perspectives for which the Internet allows can counter its democratic tendencies (cf. Sunstein [2008]). Søraker [2008], however, has argued that the increasing use of frameworks within which Internet users can contribute nontextual information constitutes a serious obstacle to government attempts to censor and monitor Internet traffic.

Whether or not the Internet has an inherent tendency to support democratic processes, it has often been argued that cyberspace ought to be organized to support direct democracy and citizenship by functioning as a public sphere, an area in social life in which people get together to discuss issues of mutual interest, and to develop shared opinions and judgments, and take political action when appropriate [Gimmler, 2001; Bohman, 2008]. Both the idea that cyberspace should function as a public sphere and that it is capable of doing so have been criticized [Dean, 2002]. The functioning of cyberspace as a public space has, in any case, come under pressure since the mid-1990s with the emergence of e-commerce and the concomitant processes of commodification and privatization. E-commerce has brought along increased governmental regulation of cyberspace, and enhanced attention to issues of intellectual property rights, security and cybercrime, as well as to commercial free speech consumer privacy, and ethical issues of pornography.

Let us now turn to philosophical issues in cyberpolitics. The central issues here is the proper use of the Internet for political purposes by governments, political parties, and interest groups. An overarching question is how the Internet can and should be used to support democratic political processes in the “real” world [Chadwick, 2006; Shane, 2004]. What should governments, providers and others do to strengthen or utilize the possibilities of the Internet for supporting “real-life” democratic politics, including better means for political communication, deliberation, participation and activism? Should all government information be made available online? Should electronic voting be introduced [Pieters, 2006]? What role can and should the Internet have in international and global political processes?

A specific political issue is raised by the *digital divide*, the existence of a gap of effective access to information technology because of preexisting imbalances in resources and skills [Norris, 2001; Van Dijk, 2005]. The digital divide has been argued to exacerbate inequalities in society, since effective access to information technology has become an important condition for social and economic success. Van den Hoven and Rooksby [2008] argue that in contemporary societies information qualifies as a Rawlsian primary good, a necessary means to realizing one's life plan, and derive a set of criteria for the just distribution of access to information. A final issue concerns the politics of national security and cyberterrorism. The Internet has become a critical infrastructure that is vulnerable to cyberattacks, and also functions where conventional terrorist attacks may be prepared and discussed. This raises the question of how much control the government should exercise over the Internet in the interest of national security [Nissenbaum, 2005] and where the boundaries should be drawn between cyberterrorism and other subversive online activities, such as cyberactivism, cybercrime and hacking [Manion and Goodrum, 2000].

5.6 Cyborgs and Virtual Subjects

Information technology has become so much part of everyday life that it is affecting human identity (understood as character). Two developments have been claimed to have a particularly great impact. The first of these is that information technologies are starting to become part of our bodies and function as prosthetic technologies that take over or augment biological functions. These technologies are changing humans into cyborgs, cybernetic organisms, and thereby altering human nature. A second development is the emergence of virtual identities, which are identities that people assume online and in virtual worlds. This development has raised questions about the nature of identity and the self, and their realization in the future.

Philosophical studies of cyborgs have considered three principal questions: the conceptual question of what a cyborg is, the interpretive and empirical question of whether humans are or are becoming cyborgs, and the normative questions of whether it would be good or desirable for humans to become cyborgs. The term "cyborg" has been used in three increasingly broad senses. The traditional definition of a cyborg, is that of an organism that is part human, part machine. Cyborgs, in this sense, are largely fictional beings that include both organic systems and artificial systems between which there is feedback-control, and in which the artificial systems closely mimic the behavior of organic systems. On a broader conception, a cyborg is any individual with artificial parts, even if these parts are simple structures like artificial teeth and breast implants. On a still broader conception, a cyborg is any individual who relies extensively on technological devices and artifacts to function. On this conception, everyone is a cyborg, since everyone relies extensively on technology.

Cyborgs have become a major research topic in cultural studies, which has brought forth the area of *cyborg theory*, which is the multidisciplinary study of cyborgs and their representation

in popular culture [Gray, 1996]. In this field the notion of the cyborg is often used as a metaphor to understand aspects of contemporary - late modern or postmodern - society's relationship to technology, as well as to the human body and the self. The advance of cyborg theory has been credited to Donna Haraway, in particular her essay "Manifesto for Cyborgs" [Haraway, 1985]. Haraway claims that the binary ways of thinking of modernity (organism-technology, man-woman, physical-nonphysical and fact-fiction) traps beings into supposedly fixed identities and oppresses those beings (animals, women, blacks, etc.) who are on the wrong, inferior side of binary oppositions. She believes that the hybridization of humans and human societies, through the notion of the cyborg, can free those who are oppressed by blurring boundaries and constructing hybrid identities that are less vulnerable to the trappings of modernistic thinking (see also Mazlish [1993]).

Haraway believes, along with many other authors in cyborg theory (cf. Gray [2004] and Hayles [1999]) that this hybridization is already occurring on a large scale. Many of our most basic concepts, such as those of human nature, the body, consciousness and reality, are shifting and taking on new, hybrid, informationalized meanings. In this postmodern, posthuman age, power relations take on new forms, and new forms of freedom and resistance are made possible. Coming from the philosophy of cognitive science Andy Clark [2003] develops the argument that technologies have always extended and co-constituted human nature (cf. Brey [2000]), and specifically human cognition. He concludes that humans are "natural-born cyborgs" (see also the discussion of Clark in Section 3.6).

Philosophers Nick Bostrom and David Pearce have founded a recent school of thought, known as *transhumanism* that shares the positive outlook on the technological transformation of human nature held by many cyborg theorists [Bostrom, 2005; Young, 2005]. Transhumanists want to move beyond humanism, which they commend for many of its values but which they fault for its belief in a fixed human nature. They aim at increasing human autonomy and happiness and eliminate suffering and pain (and possibly death) through human enhancement. Thus achieving a trans- or posthuman state in which bodily and cognitive abilities are augmented by modern technology.

Critics of transhumanism and human enhancement, like Francis Fukuyama, Leon Kass, George Annas, Jeremy Rifkin and Jürgen Habermas, oppose tinkering with human nature for the purpose of enhancement. Their position that human nature should not be altered through technology has been called *bioconservatism*. Human enhancement has been opposed for a variety of reasons, including claims that it is unnatural, undermines human dignity, erodes human equality, and can do bodily and psychological harm [DeGrazia, 2005]. Currently, there is an increasing focus on ethical analyses of specific enhancement and prosthetic technologies that are in development. Information technology-based prostheses are discussed by Gillett [2006], who considers various types of neurorehabilitative and augmentative technologies, and Lucivero and

Tamburrini [2008], who study ethical aspects of brain-computer interfaces. James Moor [2004] has cautioned that there are limitations to ethical studies of prostheses and enhancement technologies. Since ethics is determined by one's nature, he argues, a decision to change one's nature cannot be settled by ethics itself.

Questions concerning human nature and identity are also being asked anew because of the coming into existence of *virtual identities* [Maun and Corruncker, 2008]. Such virtual identities, or online identities, are social identities assumed or presented by persons in computer-mediated communication and virtual communities. They usually include textual descriptions of oneself, and can include other types of media. Virtual environments are special in that persons are represented in them by means of an *avatar*, which is a graphically realized character over which users assume control. Salient features of virtual identities are that they can be different from the corresponding real-world identities, that persons can assume multiple virtual identities in different contexts and settings, that virtual identities can be used by persons to emphasize or hide different aspects of their personality and character, and that they usually do not depend on or make reference to the user's embodiment or situatedness in real life.

In a by now classical (though also controversial) study of virtual identity, psychologist Sherry Turkle [1995] argues that the dynamics of virtual identities appear to validate poststructuralist and postmodern theories of the subject, which posit a decentered subject that stands in opposition to the Cartesian subject of modernity. The Cartesian subject is a unitary subject with a fixed and stable identity that defines who someone really is. Poststructuralist and postmodern scholars reject this essentialist conception of the self, and hold that the self is constructed, multiple, situated, and dynamical. This conception seems to make a perfect fit with virtual identity, which is similarly constructed and multiple. The next step to take is to claim that behind these different virtual identities, there is no stable self, but rather that these identities, along with other projected identities in real life, collectively constitute the subject.

The dynamics of virtual identities have been studied extensively in fields like cultural studies and new media studies. It has been mostly assessed positively that people can freely construct their virtual identities, that they can assume multiple identities in different contexts and can explore different social identities to overcome oppositions and stereotypes, that virtual identities stimulate playfulness and exploration, and that traditional social identities based on categories like gender and race play a lesser role in cyberspace [Turkle, 1995; Bell, 2001]. Critics like Dreyfus [2001] and Borgmann [1999], however, argue that virtual identities promote inauthenticity and the hiding of one's true identity, and lead to a loss of embodied presence, a lack of commitment and a shallow existence. Taking a more neutral stance, Bennan and Pettit [2008] analyze the importance of esteem on the Internet, and argue that people care about their virtual reputations even if they have multiple virtual identities. Matthews [2008], finally, considers the relation between virtual identities and cyborgs, both of which are often supported and

denounced for quite similar reasons, namely their subversion of the concept of a fixed human identity.

6. Computer and Information Ethics

This section surveys the field of computer and information ethics. The first section will define the field and will consider its aims and scope, its history, and major approaches and orientations. In the section thereafter, major topics in computer ethics will be surveyed, including privacy, security, free expression and content control, equity issues, intellectual property, and issues of moral responsibility. The final section will focus on the approaches of values in design and value-sensitive design, which aim to analyze embedded values in computer software and systems, and to devise methodologies for incorporating values into the design process.

6.1 Approaches in Computer and Information ethics

Computer ethics is a field of applied ethics that addresses ethical issues in the use, design and management of information technology and in the formulation of ethical policies for its regulation in society. For contemporary overviews of the field, see Tavani [2007], Weckert [2007], Spinello and Tavani [2004] and Himma and Tavani [2008]. Computer ethics, which has also been called *cyberethics*, took off as a field in the 1980s, together with the rise of the personal computer. Early work in the field had already started in the 1940s, soon after the invention of the computer. MIT Professor Norbert Wiener was a precursor of the field, already identifying many issues of computer ethics in his book *The Human Use of Human Beings* [Wiener, 1950]. The term "computer ethics" was first introduced in the mid-1970s by Walter Maner, who also promoted the idea of teaching computer ethics in computer science curricula [Maner, 1980]. The watershed year of 1985 saw the appearance of seminal publications by Jim Moor [1985] and Deborah Johnson [1985] that helped define the field. Since then, it has become a recognized field of applied ethics, with its own journals and conference series. In recent years, the field is sometimes also related to a more general field of *information ethics*, which includes computer ethics, media ethics, library ethics, and bioinformation ethics.

Why would there be a need for computer ethics, while there is no need for a separate field of ethics for many other technologies, like automobiles and appliances? Jim Moor [1985] has argued that the computer has had an impact like no other recent technology. The computer seems to impact every sector of society, and seems to require us to rethink many of our policies, laws and behaviors. According to Moor, this great impact is due to the fact that computers have *logical malleability*, meaning that their structure allows them to perform any activity that can be specified as a logical relation between inputs and outputs. Many activities can be specified in this way, and the computer therefore turns out to be an extremely powerful and versatile machine that

can perform an incredible amount of functions, from word processor to communication device to gaming platform to financial manager.

The versatility of computers is an important reason for the occurrence of a computer revolution, or information revolution, which is now transforming many human activities and social institutions. Many important things that humans do, including many that raise moral questions like stealing from someone, defaming someone, or invading someone's privacy now also exist in electronic form. In addition, the computer also makes substantially new types of activities possible that are morally controversial, such as the creation of virtual child pornography for which no real children were abused. Because many of the actions made possible by computers are different and new, we often lack policies and laws to guide them. They generate what Moor has called *policy vacuums*, being the lack of clear policies or rules of conduct. The task of computer ethics, then, is to propose and develop new ethical policies, ranging from explicit laws to informal guidelines, to guide new types of actions that involve computers.

Computer ethics has taken off since its birth in the mid-80s, and has established itself as a mature field with its own scientific journals, conferences and organizations. The field initially attracted most of its interests from computer scientists and philosophers, with many computer science curricula nowadays requiring a course or module on computer ethics. However, given the wide implications for human action sketched by Moor, computer ethics is also of interest to other fields that focus on human behavior and social institutions, such as law, communication studies, education, political science and management. Moreover, computer ethics is also an important topic of debate in the public arena, and computer ethicists regularly contribute to public discussions regarding the use and regulating of computer technology.

Computer ethics is sometimes defined as a branch of *professional ethics* similar to other branches like engineering ethics and journalism ethics. On this view, the aim of computer ethics is to define and analyze the moral and professional responsibilities of computer professionals. *Computer professionals* are individuals employed in the information technology branch, for example as hardware or software engineer, web designer, network or database administrator, computer science instructor or computer-repair technician. Computer ethics, on this view, should focus on the various moral issues that computer professionals encounter in their work, for instance in the design, development and maintenance of computer hardware and software.

Within this approach to computer ethics, most attention goes to the discussion of ethical dilemmas that various sorts of computer professionals may face in their work and possible ways of approaching them. Such dilemmas may include, for example, the question how one should act as a web designer when one's employer asks one to install spyware into a site built for a client, or the question to what extent software engineers should be held accountable for harm incurred by software malfunction. Next to the discussion of specific ethical dilemmas, there is also general discussion of the responsibilities of computer professionals towards various other parties, such as

clients, employers, colleagues, and the general public, and of the nature and importance of ethical codes in the profession. A recent topic of interest has been the development of methods for *value-sensitive design*, which is the design of software and systems in such a way that they conform to a desired set of (moral) values [Friedman, Kahn and Borning, 2006].

While the professional ethics view of computer ethics is important, many in the field employ a broader conception that places the focus on general ethical issues in the use and regulation of information technology. This approach may be called the *philosophical ethics* approach to computer ethics. This conception holds, following Moor [1985], that computer ethics studies moral issues that are of broad societal importance, and develops ethical policies to address them. Such policies may regulate the conduct of organizations, groups and individuals and the workings of institutions. The philosophical approach focuses on larger social issues like information privacy and security, computer crime, issues of access and equity, and the regulation of commerce and speech on the Internet. It asks what ethical principles should guide our thinking about these issues, and what specific policies (laws, social and corporate policies, social norms) should regulate conduct with respect to them. Within this approach, some researchers focus on the development of ethical guidelines for users of computer technology. Others place more emphasis on policy issues, and try to formulate ethical policies for organizations, government agencies or lawmakers. Still others focus on computer technologies themselves, and try to to identify and evaluate morally relevant features in their design. Some also focus on theoretical and metaethical issues.

6.2 Topics in Computer and Information Ethics

Introductions to computer ethics show considerable agreement on what the central issues for computer ethics are. They include ethical issues of privacy, security, computer crime, intellectual property, free expression, and equity and access, and issues of responsibility and professional ethics.

Privacy

Privacy is a topic that has received much attention in computer ethics from early on. Information technology is often used to record, store and transmit personal information, and it may happen that this information is accessed or used by third parties without the consent of the corresponding persons, thus violating their privacy. Privacy is the right of persons to control access to their personal affairs, such as their body, thoughts, private places, private conduct, and personal information about themselves. The most attention in computer ethics has gone to *information privacy*, which is the right to control the disclosure of personal data. Information technology can easily be used to violate this right.

Privacy issues come into play on the Internet, where cookies, spyware, browser tracking and access to the records of internet providers may be used to study the Internet behavior of individuals or to get access to their PCs. They also come into play in the construction of databases with personal information by corporations and government organizations, and the merging of such databases to create complex records about persons or to find matches across databases. Other topics of major concern include the privacy implications of video surveillance and biometric technologies, and the ethics of medical privacy and privacy in the workplace. It has also been studied whether people have a legitimate expectation to privacy in public areas, whether they can be freely recorded, screened and tracked whenever they appear in public and how the notion of "public" itself has changed in light of information technology.

Security and crime

Security has become a major issue in computer ethics, because of rampant computer crime and fraud, the spread of computer viruses, malware and spam, and national security concerns about the status of computer networks as breeding grounds for terrorist activity and as vulnerable targets for terrorist attacks. Computer security is the protection of computer systems against the unauthorized disclosure, manipulation, or deletion of information and against denial of service attacks. Breaches of computer security may cause harms and rights violations, including economic losses, personal injury and death, which may occur in so-called safety-critical systems, and violations of privacy and intellectual property rights.

Much attention goes to the moral and social evaluation of computer crime and other forms of disruptive behavior, including *hacking* (non-malicious break-ins into systems and networks), *cracking* (malicious break-ins), *cybervandalism* (disrupting the operations of computer networks or corrupting data), *software piracy* (the illegal reproduction or dissemination of proprietary software), and *computer fraud* (the deception for personal gain in online business transactions by assuming a false online identity or by altering or misrepresenting data). Another recently important security-related issue is how state interests in monitoring and controlling information infrastructures to better protect against terrorist attacks should be balanced against the right to privacy and other civil rights [Nissenbaum, 2005].

Free expression and content control

The Internet has become a very important medium for the expression of information and ideas. This has raised questions about whether there should be content control or censorship of Internet information, for example by governments or service providers. Censorship could thwart the right to free expression, which is held to be a basic right in many nations. Free expression includes both freedom of speech (the freedom to express oneself through publication and dissemination) and freedom of access to information.

Several types of speech have been proposed as candidates for censorship. These include pornography and other obscene forms of speech, hate speech such as websites of fascist and racist organizations, speech that can cause harm or undermine the state, such as information as to how to build bombs, speech that violates privacy or confidentiality, and libelous and defamatory speech. Studies in computer ethics focus on the permissibility of these types of speech, and on the ethical aspects of different censorship methods, such as legal prohibitions and software filters (see also Section 5.5).

Equity and access

The information revolution has been claimed to exacerbate inequalities in society, such as racial, class and gender inequalities, and to create a new, digital divide, in which those that have the skills and opportunities to use information technology effectively reap the benefits while others are left behind. In computer ethics, it is studied how both the design of information technologies and their embedding in society could increase inequalities, and how ethical policies may be developed that result in a fairer and more just distribution of their benefits and disadvantages. This research includes ethical analyses of the accessibility of computer systems and services for various social groups, studies of social biases in software and systems design, normative studies of education in the use of computers, and ethical studies of the digital gap between industrialized and developing countries.

Intellectual property

Intellectual property is the name for information, ideas, works of art and other creations of the mind for which the creator has an established proprietary right of use. Intellectual property laws exist to protect creative works by ensuring that only the creators benefit from marketing them or making them available, be they individuals or corporations. Intellectual property rights for software and digital information have generated much controversy. There are those who want to ensure strict control of creators over their digital products, whereas others emphasize the importance of maintaining a strong public domain in cyberspace, and argue for unrestricted access to electronic information and for the permissibility of copying proprietary software. In computer ethics, the ethical and philosophical aspects of these disputes are analyzed, and policy proposals are made for the regulation of digital intellectual property in its different forms. Patentability of software is also a topic of major concern, which is problematic due to the non-tangible nature of software as well as the difficulty in specifying what counts as the identity of a piece of software (cf. Turner and Eden, forthcoming b).

Moral Responsibility

Society strongly relies on computers. It relies on them for correct information, for collaboration and social interaction, for aid in decision-making, and for the monitoring and execution of tasks. When computer systems malfunction or make mistakes, harm can be done, in terms of loss of time, money, property, opportunities, or even life and limb. Who is responsible for such harms? Computer professionals, end-users, employers, policy makers and others could all be held responsible for particular harms. It has even been argued that intelligent computer systems can bear moral responsibility themselves [Dodig-Crnkovic and Persson, 2008]. In computer ethics, it is studied how the moral responsibility of different actors can be defined, and what kinds of decisions should be delegated to computers to begin with. It is studied how a proper assignment of responsibility can minimize harm and allows for attributions of accountability and liability.

Foundational Issues in Computer Ethics

Foundational, metaethical and methodological issues have received considerable attention in computer ethics. Many of these issues have been discussed in the context of the so-called *foundationalist debate* in computer ethics [Floridi and Sanders, 2002; Himma, 2007a]. This is an ongoing metatheoretical debate on the nature and justification of computer ethics and its relation to metaethical theories. Three questions central to the foundationalist debate are: "Is computer ethics a legitimate field of applied ethics?", "Does computer ethics raise any ethical issues that are new or unique?" and "Does computer ethics require substantially new ethical theories, concepts or methodologies different from those used elsewhere in applied ethics?".

The first question, whether computer ethics is a legitimate field of applied ethics, has often been discussed in the context of the other two questions, with discussants arguing that the legitimacy of computer ethics depends on the existence of unique ethical issues or questions in relation to computer technology. The debate on whether such issues exist has been called the *uniqueness debate* [Tavani, 2002]. In defense of uniqueness, Maner [1996] has argued that unique features of computer systems, like logical malleability, superhuman complexity and the ability to make exact copies, raise unique ethical issues to which no non-computer analogues exist. Others remain unconvinced that any computer ethics issue is genuinely unique. Johnson [2003] has proposed that issues in computer ethics are new species of traditional moral issues. They are familiar in that they involve traditional ethical concepts and principles like privacy, responsibility, harm and ownership, but the application of these concepts and principles is not straightforward because of special properties of computer technology, which require a rethinking and retooling of ethical notions and new ways of applying them.

Floridi and Sanders [2002; Floridi, 2003] do not propose the existence of unique ethical issues but rather argue for the need of new ethical theory. They argue that computer ethics needs a metaethical and macrotheoretical foundation, which they argue be different from the standard macroethical theories like utilitarianism and Kantianism. Instead, they propose a

macroethical theory they call Information Ethics, which assigns intrinsic value to information. The theory covers not just digital or analogue information, but in fact analyzes all of reality as having an informational ontology, being built out of informational objects. Since informational objects are postulated to have intrinsic value, moral consideration should be given to them, including the informational objects produced by computers. In contrast to these various authors, Himma [2007a] has argued that computer technology does not need to raise new ethical issues or require new ethical theories to be a legitimate field of applied ethics. He argues that issues in computer ethics may not be unique and may be approached with traditional ethical theories, and that it nevertheless is a legitimate field because computer technology has given rise to an identifiable cluster of moral issues in much the same way like medical ethics and other fields of applied ethics.

Largely separately from the foundationalist debate, several authors have discussed the issue of proper methodology in computer ethics, discussing standard methods of applied ethics and their limitations for computer ethics [Van den Hoven, 2008; Brey, 2000]. An important recent development that has methodological and perhaps also metaethical ramifications is the increased focus on cultural issues. In *intercultural information ethics* [Ess and Hongladarom, 2007; Brey, 2007], ethicists attempt to compare and come to grips with the vastly different moral attitudes and behaviors that exist towards information and information technology in different cultures. In line with this development, Gorniak-Kocykowska [1995] and others have argued that the global character of cyberspace requires a *global ethics* which transcends cultural differences in value systems.

Other Topics

There are many other social and ethical issues that are studied in computer ethics next to these central ones. Some of these include the implications of IT for community, identity, the quality of work, and the quality of life, the relation between information technology and democracy, the ethics of Internet governance and electronic commerce, and the ethics of trust online. Many new ethical issues come up together with the development of new technologies or applications. Recently, much attention has been devoted to ethical aspects of social networking sites like Facebook, MySpace and Youtube, to ubiquitous computing and ambient intelligence, and to robotics and artificial agents. The constant addition of new products and services in information technology and the emergence of new uses and correlated social and cultural consequences ensures that the field keeps meeting new challenges.

6.3 Values and Computer Systems Design

Although most ethical commentary in the philosophical approach is directed at the *use* of computers by individuals and organizations, attention has also started to be paid to systems and

software *themselves*. It has come to be recognized that the systems themselves are not morally neutral but contain values and biases in their design that must also be analyzed. Approaches of this sort have been called *values in design* approaches [Nissenbaum, 1998; Flanagan, Howe and Nissenbaum, 2007]. Values in design approaches hold that computer software and systems can be morally evaluated partially or wholly independently of actual uses of them. They can be said to embody values in the sense that they have a tendency to promote or sustain particular values when used.

This may sound like technological determinism, but proponents usually do not subscribe to the strong determinist thesis that embodied values necessitate certain effects in whatever way the system is used. Yet, they do hold a weak determinism according to which systems may embody values that systematically engender certain effects across a wide range of uses, at least including typical or "normal" ways of using the system. For a system to embody a value, then, means that there is a tendency for that value to be promoted or realized when the system is used. This observation has led proponents to argue that more attention should be paid to ethical aspects in the design of computer systems rather than just their use.

Friedman and Nissenbaum [1996] have studied how values may enter into computer systems, with a focus on justice and bias. They argue that bias can enter into computer systems in three ways. Preexisting bias emerges from the practices and attitudes of designers and the social institutions in which they function. Technical bias arises from technical constraints. Emergent bias arises after the design of the system, when a context of use emerges that is different from the one anticipated. These three origins of bias may be generalized to apply to other values as well.

Brey [2000] has proposed a particular values in design approach termed *disclosive computer ethics*. He claims that a significant part of the effort of computer ethics should be directed at deciphering and subsequently evaluating embedded moral values in computer software and systems. The focus should be on widely held public and moral values, such as privacy, autonomy, justice, and democracy. Research, Brey argues, should take place at three levels: the disclosure level, at which morally charged features of computer systems are detected and disclosed, the theoretical level, at which relevant moral theory is developed, and the application level, at which ethical theory is used in the evaluation of the disclosed morally charged features. He claims that such research should be interdisciplinary, involving ethicists, computer scientists and social scientists.

The approach of *value-sensitive design* [Friedman, Kahn and Borning, 2006; Friedman & Kahn, 2003] is not so much concerned with the identification and evaluation of values in computer systems, but rather with the development of methods for incorporating values into the design process. It is an approach to software engineering and systems development that builds on values in design approaches and studies how accepted moral values can be operationalized and

incorporated into software and systems. Its proposed methodology integrates conceptual investigations into values, empirical investigations into the practices, beliefs and intentions of users and designers, and technical investigations into the way in which technological properties and mechanisms support or hinder the realization of values. It also seeks procedures to incorporate and balance the values of different stakeholders in the design process.

Acknowledgments

The authors would like to thank the following individuals for their helpful comments and recommendations: Adam Briggle, Terrell Ward Bynum, Andy Clark, Gordana Dodig-Crnkovic, Amnon H. Eden, Timothy Colburn, Juan Manuel Duran, Charles Ess, James H. Fetzer, Sven Ove Hansson, David Harel, Luciano Floridi, Patrick Grim, David Koepsell, Mauriece Liebregt, Anthonie Meijers, William J. Rapaport, Oron Shagrir, Herman Tavani, Raymond Turner and Alasdair Urquhart. We have also benefited from feedback given by participants at the E-CAP ’08 conference at Montpellier, France.

Bibliography

P. Abrahams. What Is Computer Science? *Communications of the ACM*, 30(6), 472-473, 1987

A. Abran, J.W. Moore, P. Bourque and R. Dupuis. *SWEBOK – Guide to the Software Engineering Body of Knowledge*.IEEE Computer Society Press, 2004

P. Adriaans and J. van Benthem, eds. *Handbook on the Philosophy of Information*. Elsevier, forthcoming

C. Allen, I. Smith and W. Wallach. Artificial morality: Top-down, bottom-up, and hybrid approaches, *Ethics and Information Technology,* 7, 149-155, 2005

M.L. Anderson. Embodied Cognition: A field guide. *Artificial Intelligence*, 149(1), 91-130, 2003

P. M. Asaro. How Just Could a Robot War be? In *Current Issues in Computing and Philosophy*, K. Waelbers, A. Briggle and P. Brey, eds. IOS Press, forthcoming

I. Asimov. *I, Robot*. Grafton Books, 1968

Y. Bar-Hillel. *Language and Information: Selected Essays on Their Theory and Application. A*ddison-Wesley, 1964.

J.A. Barker. Computer Modeling and the Fate of Folk Psychology. In *Cyberphilosophy*, J.H. Moor and T.W. Bynum, eds., pp. 26-44. Blackwell, 2002.

D. Barker-Plummer . Turing Machines. In *The Stanford Encyclopedia of Philosophy (Winter 2007 Edition*), E.N. Zalta, ed., 2007

D. Barney. The Vanishing Table, or Community in a World That is No World. In *Community in the Digital Age: Philosophy and Practice*, A. Feenberg and D. Barney, eds., pp. 31-52. Rowman and Littlefield, 2004

J. Baudrillard. *Simulacra and Simulation.* University of Michigan Press, 1995

W. Bechtel and A. Abrahamson. *Connectionism and the Mind: An Introduction to Parallel Processing in Networks*. Blackwell, 1990

M. A. Bedau. Artificial Life: Organization, adaptation and complexity from the bottom up, *Trends in Cognitive Sciences*, 7(11), 505-512, 2003

M. A. Bedau. Artificial Life. In *The Blackwell Guide to Philosophy of Computing and Information*, L. Floridi, ed., pp. 197-212, Blackwell, 2004

M. A. Bedau. Philosophical Content and Method of Artificial Life. In *The Digital Phoenix: How Computers are Changing Philosophy*, T. W. Bynum and J. H. Moor, eds., pp. 135-152. Blackwell, 1998

D. Bell. *An Introduction to Cybercultures*. Routledge, 2001

M. Benedikt, ed. *Cyberspace: First Steps*. MIT Press 1991

W. Benjamin. The Work of Art in the Age of Mechanical Reproduction. In *Media and Cultural Studies: Key works*, M. G. Durham & D. Kellner, eds., pp. 18-40. Oxford: Blackwell, 2006

P. Blackburn and J. Bos. *Representation and Inference for Natural Language*. Chicago University Press, 2005

M. Boden, ed. *The Philosophy of Artificial Intelligence*. Oxford University Press 1990

M. Boden, ed. *The Philosophy of Artificial Life*. Oxford University Press, 1996

M. Boden. *Mind As Machine: A History of Cognitive Science.* Oxford University Press 2006

J. Bohman. The Transformation of the Public Sphere: Political Authority, Communicative Freedom, and Internet Publics. In *Information Technology and Moral Philosophy*, J. van den Hoven and J. Weckert, eds., pp. 66-92. Cambridge University Press, 2008

A. Borgmann. *Holding On to Reality: The Nature of Information at the Turn of the Millennium*. University Of Chicago Press, 1999

A. Borgmann. Is the Internet the Solution to the Problem of Community? In *Community in the Digital Age: Philosophy and Practice*, A. Feenberg and D. Barney, eds., pp. 53-68. Rowman and Littlefield, 2004

G. Brennan and P. Pettit. Esteem, Identifiability, and the Internet. In *Information Technology and Moral Philosophy*, J. van den Hoven and J. Weckert, eds., pp. 175-194. Cambridge University Press, 2008

P. Brey. New Media and the Quality of Life. *Techné: Journal of the Society for Philosophy and Technology* 3(1), 1-23, 1998a

P. Brey. Space-Shaping Technologies and the Geographical Disembedding of Place, In *Philosophy & Geography vol. III: Philosophies of Place,* A. Light and B. & J. B. Smith, pp. 239-263. Rowman & Littlefield, 1998b

P. Brey. Disclosive Computer Ethics: The Exposure and Evaluation of Embedded Normativity in Computer Technology, *Computers and Society*, 30(4), 10-16, 2000.

P. Brey. Technology as Extension of Human Faculties, In *Metaphysics, Epistemology, and Technology. Research in Philosophy and Technology, vol 19*, C. Mitcham, ed. Elsevier/JAI Press, 2000

P. Brey. The Social Ontology of Virtual Environments. In *John Searle's Ideas About Social Reality: Extensions, Criticisms and Reconstructions,* D. Koepsell and L. Moss, eds., pp. 269-282. Blackwell 2003

P. Brey. The Epistemology and Ontology of Human-Computer Interaction. *Minds and Machines 15*(3-4), 383-398, 2005

P. Brey. Is Information Ethics Culture-Relative? *International Journal of Technology and Human Interaction.* Special Issue *Information Ethics: Global Issues, Local Perspectives*. 3(3), 12-24, 2007

P. Brey. Virtual Reality and Computer Simulation. In *The Handbook of Information and Computer Ethics,* K. Himma and H. Tavani, eds., John Wiley and Sons: 2008

R. A. Brooks. *Flesh and Machines, How Robots Will Change Us.* Pantheon Books, 2002

J. G. Brookshear. *Computer Science: an overview [10th ed.].* Addison-Wesley, 2007

L. Burkholder, ed. *Philosophy and the Computer*. Westview Press, 1992

T. Butler and C. Murphy. Understanding the design of information technologies for knowledge management in organizations: a pragmatic perspective. *Information Systems Journal* 17(2), 143-163, 2007

T. W. Bynum and J. Moor. How Computers are Changing Philosophy. In *The Digital Phoenix: How Computers are Changing Philosophy*. T. Bynum and J. Moor, eds., pp. 1-16. Blackwell, 1998a

T. W. Bynum and J. Moor, eds. *The Digital Phoenix: How Computers are Changing Philosophy*. Blackwell, 1998b

M. Castells. *The Rise of the Network Society. Information Age vol. 1*. Blackwell, 1996

A. Chadwick. *Internet Politics: States, Citizens, and New Communication Technologies*. Oxford University Press, 2006

P. M. Churchland. *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science*. MIT Press, 1992

A. Clark. *Microcognition: Philosophy, Cognitive Science, and Parallel Distributed Processing.* MIT Press, 1991

A. Clark. *Associative Engines: Connectionism, Concepts, and Representational Change*. MIT Press, 1993

A. Clark. *Being there: putting brain, body, and world together again*, MIT Press, 1997

A. Clark. *Natural-born cyborgs: minds, technologies, and the future of human intelligence*. Oxford University Press, 2003

A. Clark and D. Chalmers. The Extended Mind. *Analysis* 58 (1): 7-19, 1998

R. Clarke. Asimov's laws of robotics: Implications for information technology – Part 2. *Computer*, 27(1), 57-66, 1994

D. Cocking. Plural Selves and Relational Identity: Intimacy and Privacy Online. In *Information Technology and Moral Philosophy*, J. van den Hoven and J. Weckert, eds., pp. 123-141. Cambridge University Press, 2008

T. Colburn. *Philosophy and Computer Science*. Sharpe, 1999

T. Colburn. Methodology of Computer Science. In *The Blackwell Guide to Philosophy of Computing and Information*, L. Floridi, ed., pp. 318–326, Blackwell, 2004

T. Colburn and G. Shute. Abstraction in Computer Science. *Minds and Machines 17(2), 169-184, 2007*

C. Cope, P. Horan and M. Garner. Conceptions of an Information System and Their Use in Teaching about IS. *Informing science* 1(1), 9-22, 1997

B.J. Copeland. *Artificial Intelligence: A Philosophical Introduction*. Wiley-Blackwell, 1993

B.J. Copeland. Hypercomputation, *Minds and Machines*, 12(4), 461-502, 2002a

B.J. Copeland. The Church-Turing Thesis. In *The Stanford Encyclopedia of Philosophy (Fall 2002 Edition*), E.N. Zalta, ed., 2002b

B.J. Copeland. Computation. In *The Blackwell Guide to Philosophy of Computing and Information*, L. Floridi, ed., pp. 3–17, Blackwell, 2004

B.J. Copeland and O. Shagrir. Physical Computation: How General are Gandy's Principles for Mechanisms? *Minds and Machines* 17(2), 217-231, 2007

R. Cordeschi. Cybernetics. In *The Blackwell Guide to Philosophy of Computing and Information*, L. Floridi, ed., pp. 186-196, Blackwell, 2004

P. Danielson. How Computers Extend Artificial Morality. In *The Digital Phoenix: How Computers are Changing Philosophy*. T. Bynum and J. Moor, eds., pp. 292-307. Blackwell, 1998

J. Dean. Why the Net is not a public sphere. *Constellations*, 10(1), 95-112, 2003

D. DeGrazia. Enhancement Technologies and Human Identity. *Journal of Medicine and Philosophy*, 30, 261-283, 2005

D. Dennett. Artificial Life as Philosophy. *Artificial Life* 1(3), 291-292, 1994

J. van Dijk. *The Deepening Divide: Inequality in the Information Society*. Sage, 2005

J. van Dijk. *The Network Society*. Sage, 2006

E. Dijkstra. Go To Statement Considered Harmful. *Communications of the ACM*, 11(3), 147–148, 1968

P. J. Dobson Critical realism and information systems research: why bother with philosophy? *Information Research*, 7(2) , Rec. No, 124, 2002

G. Dodig-Crnkovic. *Investigations into Information Semantics and Ethics of Computing*, Mälardalen University Press, 2006

G Dodig-Crnkovic and D. Persson. Sharing Moral Responsibility with Robots: A Pragmatic Approach. In *Tenth Scandinavian Conference on Artificial Intelligence SCAI 2008. Volume 173 of Frontiers in Artificial Intelligence and Applications*, A. Holst, P. Kreuger and P. Funk, eds., pp. 165-168. IOS Press, 2008

P. Dourish. *Where the Action is*. MIT Press, 2001

F. Dretske. *Knowledge and the Flow of Information*. MIT Press, 1981

H. L. Dreyfus. *What Computers Can't Do: A Critique of Artificial Reason.* Harper and Row, 1972

H. L. Dreyfus. *What Computers Still Can't Do: A Critique of Artificial Reason*. MIT Press, 1992

H. L. Dreyfus. Anonymity versus commitment: the dangers of education on the internet. *Ethics and information technology,* 1, 15-21, 1999

H. L. Dreyfus. *On the Internet.* Routledge, 2001

H. L. Dreyfus. Nihilism on the Information Highway: Anonymity versus Commitment in the Present Age. In *Community in the Digital Age: Philosophy and Practice*, A. Feenberg and D. Barney, eds., pp. 69-82. Rowman and Littlefield, 2004

H. L. Dreyfus and S. E. Dreyfus. From Socrates to Expert Systems: The Limits and Dangers of Calculative Rationality. *Technology in Society, 6(3)*, 217-233, 1984

H. L. Dreyfus and S. E. Dreyfus. *Mind over Machine: The Power of Human Intuition and Expertise in the Era of the Computer*. Free Press, 1986

H. L. Dreyfus and S. E. Dreyfus. Making a Mind Versus Modeling the Brain: Artificial Intelligence Back at a Branchpoint, *Daedalus,* 117, 15-43, 1988

A. Eden. Three Paradigms of Computer Science. *Minds & Machines*, 17,135–167, 2007

A. E. Eiben and J. E. Smith. *Introduction to Evolutionary Computing.* Springer, 2003

J. L. Elman. Connectionism, artificial life, and dynamical systems: New approaches to old questions. In *A Companion to Cognitive Science,* W. Bechtel and G. Graham, eds., pp. 488-505. Basil Blackwood, 1998

C. Ess, ed. *Philosophical Perspectives on Computer-Mediated Communication*. SUNY Press, 1996

C. Ess. Cultures in collision: Philosophical lessons from computer-mediated communication, *Metaphilosophy, 33 (1/2),* 229-253, 2002

C. Ess. Computer-Mediated Communication and Human-Computer Interaction. In *The Blackwell Guide to Philosophy of Computing and Information*, L. Floridi, ed., pp. 76-91, Blackwell, 2004

C. Ess and S. Hongladarom, *Information Technology Ethics: Cultural Perspectives*. IGI Global, 2007

C. Ess and F. Sudweeks, eds. Special Theme: Culture and Computer-Mediated Communication. *Journal of Computer-Mediated Communication*, 11(1), 179-394, 2005

D. Fallis. On Verifying the Accuracy of Information: Philosophical Perspectives. *Library Trends*, 52(3), 463-487, 2004

A. Feenberg and D. Barney. *Community in the Digital Age. Philosophy and Practice.* Rowman & Littlefield, 2004

J.H. Fetzer. Program Verification: The Very Idea. *Communications of the ACM*, 31(9), 1048-1063, 1988

J.H. Fetzer. Philosophical aspects of program verification. *Minds and Machines* 1(2), 197-216, 1991

J.H. Fetzer. Philosophy and Computer Science: Reflections on the Program Verification Debate. In *The Digital Phoenix: How Computers are Changing Philosophy*, T. W. Bynum and J. H. Moor, eds., pp. 253-273. Blackwell, 1998

J.H. Fetzer. The Philosophy of AI and Its Critique. In *The Blackwell Guide to Philosophy of Computing and Information*, L. Floridi, ed., pp. 119-134, Blackwell, 2004

J. Fielding, J. Simon, W. Ceusters and B. Smith. Ontological Theory for Ontological Engineering: Biomedical Systems Information Integration. *Proceedings of the Ninth International Conference on the Principles of Knowledge Representation and Reasoning*, 2004

M. Flanagan, D. Howe and H. Nissenbaum. Embodying Values in Technology: Theory and Practice. In *Information Technology and Moral Philosophy*, J. Van den Hoven and J. Weckert, eds., pp. 322-353. Cambridge University Press, 2008

T. Flew. *New Media: an Introduction*, Oxford University Press, 2002

L. Floridi. *Philosophy and Computing*. Routledge 1999a

L. Floridi. Information Ethics: On the Theoretical Foundations of Computer Ethics. *Ethics and Information Technology*,1(1), 37-56, 1999b

L. Floridi. What is the Philosophy of Information? *Metaphilosophy*, 33(1/2), 123-145, 2002

L. Floridi. On the Intrinsic Value of Information Objects and the Infosphere. *Ethics and Information Technology*, 4(4), 287-304, 2003

L. Floridi. Open Problems in the Philosophy of Information, *Metaphilosophy*, 35(4), 554-582, 2004a

L. Floridi. Information. In *The Blackwell Guide to Philosophy of Computing and Information*, L. Floridi, ed., pp. 40-62, Blackwell, 2004b

L. Floridi, ed. *The Blackwell Guide to Philosophy of Computing and Information*, Blackwell, 2004c

L. Floridi. Semantic Conceptions of Information. In *The Stanford Encyclopedia of Philosophy* (Spring 2007 Edition), E.N. Zalta, ed., 2007

L. Floridi. Peering into the Future of the Infosphere. *TidBITS,* Sep 25, 2006 [Retrieved on May 13, 2008, from http://db.tidbits.com/article/8686], 2006

L. Floridi and J. W. Sanders. Mapping the Foundationalist Debate in Computer Ethics. *Ethics and Information Technology*, 4(1), 1-9, 2002

L. Floridi and J. W. Sanders. On the Morality of Artificial Agents. *Minds and Machines*, 14(3), 349-379, 2004

R. Floyd. The paradigms of programming. *Communications of the ACM,* 22(8), 455-160, 1979

J. Fodor. *The Language of Thought.* Harvard University Press, 1975

J. Fodor and Z. Pylyshyn. Connectionism and Cognitive Architecture: a Critical Analysis, *Cognition*, 28, 3-71, 1988

E. Fredkin. An Introduction to Digital Philosophy, *International Journal of Theoretical Physics*, 42, 189-247, 2003

B. Friedman and P. H. Kahn. Human values, ethics, and design. In *The human-computer interaction handbook,* J. A. Jacko and A. Sears, eds., pp. 1177-1201. Lawrence Erlbaum Associates, 2003

B. Friedman and H. Nissenbaum. Bias in Computer Systems, *ACM transactions on information systems*, 14 (3), 330 – 347, 1996

B. Friedman, P. H. Kahn and A. Borning. Value Sensitive Design and Information Systems. In *Human-Computer Interaction in Management Information Systems: Foundations,* P. Zhang and D. Galletta, eds., pp. 348-372. M.E. Sharpe, 2006

R. Frigg and J. Reiss: The Philosophy of Simulation: Hot New Issues or Same Old Stew? *Synthese*, forthcoming

A. Galton. The Church–Turing thesis: Still valid after all these years? *Applied Mathematics and Computation*, 178, 93–102, 2006

T. van Gelder. The dynamical hypothesis in cognitive science. *Behavioral and Brain Sciences*, 21, 615-628, 2000

G. Gillett. Cyborgs and moral identity. *Journal of Medical Ethics,* 32, 79-83, 2006

D. Gillies. Probability in Artificial Intelligence. In *The Blackwell Guide to Philosophy of Computing and Information*, L. Floridi, ed., pp. 276-288, Blackwell, 2004

A. Gimmler. Deliberative democracy, the public sphere and the internet. *Philosophy and Social Criticism,* 27(4), 21-39, 2001

J. Gips. Toward the ethical robot. In *Android Epistemology*, K. M. Ford, C. Glymour and P. Hayes, eds. MIT Press, 1994

A. Goldman. *Epistemology and cognition*. Harvard University Press, 1986

A. Goldman. *Knowledge in a Social World*. Oxford University Press, 1999

A. Goldman. The Social Epistemology of Blogging. In *Information Technology and Moral Philosophy*, J. van den Hoven and J. Weckert, eds., pp. 111-122. Cambridge University Press, 2008

K. Gorniak-Kocikowska. The Computer Revolution and the Problem of Global Ethics. *Science and Engineering Ethics*, 2(2), 177-190, 1995

P. Graham. *Hackers and Painters: Big Ideas from the Computer Age*. O'Reilly and Associates, 2004

C. Gray, ed. *The Cyborg Handbook*. Routledge, 1996

P. Grim. Computational Modeling as a Philosophical Methodology. In *The Blackwell Guide to Philosophy of Computing and Information*, L. Floridi, ed., pp. 337–347, Blackwell, 2004

P. Grim. Philosophy for computers: Some explorations in philosophical modeling. In *Cyberphilosophy*, J.H. Moor and T.W. Bynum, eds., pp. 173-200. Blackwell, 2002.

P. Grim, G. Mar and P. St. Denis. *The Philosophical Computer: Exploratory Essays in Philosophical Computer Modeling*. MIT Press, 1988

P. Grim, E. Selinger, W. Braynen, R. Rosenberger, R. Au, N. Louie, and J. Connolly. Modeling Prejudice Reduction. *Public Affairs Quarterly* 19(2), 95-125, 2005

T.R. Gruber. Toward Principles for the Design of Ontologies Used for Knowledge Sharing. *International Journal of Human-Computer Studies*, 43(5-6), p. 907-928, 1995

N. Guarino, ed. *Formal Ontology in Information Systems*. IOS Press, 1998

A. Hagar. Quantum Algorithms: Philosophical Lessons. *Minds and Machines*, 17(2), 233 – 247, 2007

D. Haraway. A Manifesto for Cyborgs: Science, Technology, and Socialist Feminism in the 1980's, *Socialist Review,* 80, 56-107, 1985

D. Harel. *Computers Ltd. What they* really *can't do*. Oxford University Press, 2000

D. Harel and Y. A. Feldman. *Algorithmics: the spirit of computing [3rd ed.]*. Pearson, 2004

S. Harnad. The Symbol Grounding Problem. *Physica D,* 42, 335-346, 1990

S. Harnad. Artificial life: Synthetic vs. virtual. In *Artificial Life III*, C. Langton, ed., pp. 539-552. Addison-Wesley, 1994

D. Harvey. *The Condition of Postmodernity: An Enquiry into the Origins of Cultural Change*. Blackwell, 1989

W. F. G. Haselager. Robotics, philosophy and the problems of autonomy. In *Cognitive Technologies and the Pragmatics of Cognition*, I. Dror, ed, pp. 61-77. John Benjamins, 2007

J. Haugeland. The Nature and Plausibility of Cognitivism. *Behavioral and Brain Sciences,* 2, 215-260, 1978

J. Haugeland, ed. *Mind Design: Philosophy, Psychology, and Artificial Intelligence*. MIT Press, 1981

P. Hayes. The Naïve Physics Manifesto. In *The Philosophy of Artificial Intelligence*. M. Boden, ed., pp. 171-205. Oxford University Press, 1990

K. Hayles *How We Became Posthuman: Virtual Bodies in Cybernetics, Literature and Informatics*, University Of Chicago Press, 1999

M. Heim. *The Metaphysics of Virtual Reality*. Oxford University Press, 1993

D. Hilbert and W. Ackermann. *Grundzüge der theoretischen Logik.* Springer, 1928

K. E. Himma. Foundational Issues in Information Ethics. *Library Hi Tech*, 25(1), 79-94, 2007a

K. E. Himma. The concept of information overload: A preliminary step in understanding the nature of a harmful information-related condition. *Ethics and Information Technology*, 9(4), 259-272, 2007b

K. Himma and H. Tavani, eds. *Information and Computer Ethics*. John Wiley and Sons, 2008

C. Hoare. Mathematics of programming, *Byte,* August 1986, 115-49, 1986

J. Hollan, E. Hutchins and D. Kirsh. Distributed Cognition: Toward a New Foundation for Human-Computer Interaction Research', *ACM Transactions on Computer-Human Interaction,* 7(2), 174–196, 2000

J. van den Hoven. Moral Methodology and Information Technology. In *The Handbook of Information and Computer Ethics*. K. E. Himma and H. T. Tavani, eds., pp. 49-68. John Wiley and Sons, 2008J. van den Hoven and E. Rooksby. Distributive Justice and the Value of Information: A (Broadly) Rawlsian Approach. In *Information Technology and Moral Philosophy*, J. van den Hoven and J. Weckert, eds., pp. 376-398. Cambridge University Press, 2008

P. Humphreys: The Philosophical Novelty of Computer Simulation Methods. *Synthese*, forthcoming

E. Hutchins. *Cognition in the wild*, MIT Press, 1995

N. Immerman. Computability and Complexity. In *The Stanford Encyclopedia of Philosophy (Fall 2006 Edition*), E.N. Zalta, ed., 2006

L. Introna. *Management, information and power: A narrative of the involved manager*. Macmillan, 1997

D. Johnson. *Computer Ethics* [1st ed]. Prentice Hall, 1985

D. Johnson. Is the global information infrastructure a democratic technology? *Computers and Society*, 27(3), 20 – 26, 1997

D. Johnson. Computer Ethics. In *Blackwell Companion to Applied Ethics,* R.G. Frey and C.H. Wellman, eds., pp. 608-619. Blackwell, 2003

D. Johnson. Computer systems: Moral entities but not moral agents. *Ethics and Information Technology,* 8, 195-204, 2006

S. Kendal and M. Creen. *An Introduction to Knowledge Engineering*. Springer, 2006

D. Knuth, Computer Science and Its Relation to Mathematics, *American Mathematical Monthly,* 81(4), 323-343, 1974a

D. Knuth. Computer Programming as an Art. *Communications of the ACM*, 17(12), 667–673, 1974b

D. R. Koepsell. *The Ontology of Cyberspace: Law, Philosophy, and the Future of Intellectual Property*. Open Court, 2000

T. S. Kuhn. *The Structure of Scientific Revolutions [2nd ed.].* University of Chicago Press, 1970

P. de Laat. Trusting Virtual Trust. *Ethics and Information Technology,* 7(3), 167 – 180, 2005

N. J. Lacey and M. H. Lee. The Epistemological Foundations of Artificial Agents. *Minds and Machines,* 13, 339–365, 2003

C. Langton. *Artificial Life: An Overview*. MIT Press, 1995

D. Lenat and R. V. Guha. *Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project*. Addison-Wesley, 1990

D. Levy. Information Overload. In *The Handbook of Information and Computer Ethics,* K. Himma and H. Tavani, eds. John Wiley and Sons: 2008

S. Lloyd. *Programming the Universe: A Quantum Computer Scientist Takes on the Cosmos*, Alfred A. Knopf, 2006

G. Longo, ed. *Philosophy of Computer Science*, special issue in *The Monist* 82(1), 1999.

M.C. Loui. Computer Science Is an Engineering Discipline. *Engineering Education*, 78(3), 175-178, 1987

F. Lucivero and G. Tamburrini. Ethical monitoring of brain-machine interfaces. *AI & Society,* 22(3), 449 – 460, 2008

J. Lyotard. *The Postmodern Condition*. Manchester University Press, 1984

B.J. Maclennan. Transcending Turing Computability. *Minds and Machines*, 13(1), 3-22, 2003

M.S. Mahoney. Software as Science—Science as Software. In *Proceedings of the international Conference on History of Computing: Software Issues*, U. Hashagen, R. Keil-Slawik, and A. Norberg, eds., pp. 25-48. Springer, 2002

M.S. Mahoney. Calculation – Thinking – Computational Thinking: Seventeenth-Century Perspectives on Computational Science. In *Form, Zahl, Ordnung: Studien zur Wissenschafts- und Technikgeschichte*, I. Schneider, R. Seising, M. Folkerts and U. Hashagen, eds., pp. 209-222. Franz Steiner Verlag, 2004

D. Mainon and A. Goodrum. Terrorism or Civil Disobedience: Toward a Hacktivist Ethic. *Computers and Society*, 30(2), 14-19, 2000

W. Maner. *Starter Kit in Computer Ethics*, Helvetia Press, 1980

W. Maner. Unique Ethical Problems in Information Technology. *Science and Engineering Ethics*, 2(2), 137-154, 1996

E. Marcos and A. Marcos. A Philosophical Approach to the Concept of Data Model: Is a Data Model, in Fact, a Model? *Information Systems Frontiers*, 3(2), 267–274, 2001

S. Matthews. Identity and Information Technology. In *Information Technology and Moral Philosophy*, J. van den Hoven and J. Weckert, eds., pp. 142-160. Cambridge University Press, 2008

C. Maun and L. Corrunker, eds. *Virtual Identities: The Construction of Selves in Cyberspace*. Eastern Washington University Press, 2008

B. Mazlish. *The Fourth Discontinuity: The Co-evolution of Humans and Machines*. Yale University Press, 1993.

D. McDermott. A Critique of Pure Reason. In *The Philosophy of Artificial Intelligence*. M. Boden, ed., pp. 206-230. Oxford University Press, 1990

B. P. McLaughlin. Computationalism, Connectionism, and the Philosophy of Mind. In *The Blackwell Guide to Philosophy of Computing and Information*, L. Floridi, ed., pp. 135-152, Blackwell, 2004

M. McLuhan. The Gutenberg Galaxy: The Making of Typographic Man. *University of Toronto Press*, 1962

M. Minsky. *Computation: Finite and Infinite Machines*. Prentice-Hall, 1967

J. H. Moor. Are There Decisions Computers Should Never Make? *Nature and System*, 1, 217-229, 1979

J. H. Moor. What is Computer Ethics? *Metaphilosophy,* 16, 266-275, 1985

J. H. Moor, ed. *The Turing Test: The Elusive Standard of Artificial Intelligence.* Kluwer, 2003

J. H. Moor. Should we let computers get under our skin?" In *The Impact of the Internet on Our Moral Lives,* R. Cavalier, ed., pp. 121-137. SUNY press, 2005

J. H. Moor and T.W. Bynum, eds. *Cyberphilosophy.* Blackwell, 2002

N. Mooradian. Virtual Reality, Ontology, and Value. *Metaphilosophy,* 37(5), 673–690, 2006

M. Morgan and M. Morrison. *Models as Mediators: Perspectives on Natural and Social Science*. Cambridge University Press 1999

P. Mullins. Sacred Text in the Sea of Texts: The Bible in North American Electronic Culture. In *Philosophical Perspectives on Computer-Mediated Communication,* C. Ess (ed.). SUNY Press, 1996

A. Newell and, H.A. Simon. Computer Science as Empirical Inquiry: Symbols and Search. *Communications of the ACM*, 19(3), 113–126, 1976

H. Nissenbaum. Values in the Design of Computer Systems. *Computers and Society*, March 1998, 38-39, 1998

H. Nissenbaum. Securing Trust Online: Wisdom or Oxymoron. *Boston University Law Review,* 81(3), 635-664, 2001

H. Nissenbaum. Where Computer Security Meets National Security. *Ethics and Information Technology,* 7, 61-73, 2005

H. Nissenbaum and D. Walker. Will computers dehumanize education: A grounded approach to values at risk. *Technology in Society,* 20, 237-273, 1998

D.A. Norman. *Things that make us smart : defending human attributes in the age of the machine*, Addison-Wesley, 1993

P. Norris. *Digital Divide: Civic Engagement, Information Poverty, and the Internet Worldwide*. Cambridge University Press, 2001

W. Parker: Does Matter Really Matter? Computer Simulations, Experiments, and Materiality. *Synthese*, forthcoming

D. L. Parnas. Software Aspects of Strategic Defense Systems. *Communications of the ACM,* 28(12), 1326-1335, 1985

D. L. Parnas. Software Engineering Programs Are Not Computer Science Programs. *IEEE Software* 16(6), 19-30, 1999

R. Penrose. *Shadows of the Mind.* Oxford University Press, 1994

M. Peschl and C. Stary. The Role of Cognitive Modeling for User Interface Design Representations: An Epistemological Analysis of Knowledge Engineering in the Context of Human-Computer Interaction. *Minds and Machines*, 8(2), 203 – 236, 1998

P. Pettit. Trust, Reliance and the Internet, *Analyse & Kritik*, 26, 108-121, 2004

W. Pieters. Internet Voting: A Conceptual Challenge to Democracy.In *IFIP International Federation for Information Processing*, 208, 89-103, 2006

N. F. du Plooy. Information systems as social systems. In *Critical Reflections on information Systems: A Systemic Approach*, J. J. Cano, ed., pp. 105-121. IGI Publishing, 2003

R. F. Port and T. van Gelder. *Mind as Motion: Explorations in the Dynamics of Cognition*. MIT Press, 1995

M. Poster. *The Mode of Information: Poststructuralism and Social Context.* Polity, 1990.

J. Preston and M. Bishop, eds. *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence.* Oxford University Press 2002

B. T. Prosser and A. Ward. Kierkegaard and the Internet: Existential reflections on education and community. *Ethics and information technology,* 2, 167-180, 2000

R. Putnam. *Bowling Alone: The Collapse and Revival of American Community*. Simon & Schuster, 2001

Z. W. Pylyshyn. *Computation and Cognition*. MIT Press, 1984

Z. Pylyshyn, ed. *The robot's dilemma: The frame problem in artificial intelligence*. Ablex, 1987

W. V.O. Quine. *From a Logical Point of View*. Harvard University Press, 1961

W. Rapaport. Philosophy of Computer Science: An Introductory Course. *Teaching Philosophy,* 28(4), 319-41, 2005

H. Rheingold. *The Virtual Community: Homesteading on the Electronic Frontier*. Perseus, 1993

H. G. Rice. Classes of Recursively Enumerable Sets and Their Decision Problems. *Transactions of the American Mathematical Society*, 74(2), 358-366, 1953

D. Rowe. *The Hilbert Challenge*. Oxford University Press, 2000

M. Scheutz (ed.). *Computationalism: New directions*. MIT Press, 2002.

J. Searle. Minds, Brains and Programs. *Behavioural and Brain Sciences, 3(3)*, 417-424, 1980

J. Searle. *The Construction of Social Reality*. MIT Press,1995.

O. Shagrir. Effective Computation by Humans and Machines. *Minds and Machines,* 12(2), 221-240, 2002

O. Shagrir and I. Pitowski. Physical hypercomputation and the Church Turing thesis. *Minds and Machines*, 13(1), 87-101, 2003

P. M. Shane. *Democracy Online: The Prospects for Political Renewal Through the Internet* Routledge, 2004

C. Shannon. *Collected Papers*. IEEE Press, 1993

S. C. Shapiro. Computationalism. *Minds and Machines*, 5 (4), 467-87, 1995

D. Shenk. *Data Smog: Surviving the Information Glut* [revised ed.]. HarperOne, 1998

H. A. Simon. *The Shape of Automation for Men and Management*. Harper & Row, 1965

H. A. Simon. *The Sciences of the Artificial [3rd ed.]*. MIT Press, 1996

A. Sloman. *The Computer Revolution in Philosophy*. Harvester Press, 1978

A. Sloman. The irrelevance of Turing machines to Artificial Intelligence. In *Computationalism: New directions*, M. Scheutz, ed., pp. 87-128. MIT Press, 2002

B. C. Smith. Limits of Correctness in Computers. In *Computerization and Controversy*, R. Kling, ed., pp. 810-825. Academic Press, 1996

B. Smith. Ontology. In *The Blackwell Guide to Philosophy of Computing and Information*, L. Floridi, ed., pp. 155-166, Blackwell, 2004

B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, The OBI Consortium, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S.Sansone, R. H. Scheuermann, N. Shah, P. L. Whetzel and S. Lewis. The OBO Foundry: Coordinated Evolution of Ontologies to Support Biomedical Data Integration. *Nature Biotechnology*, 25(11), 1251-1255, 2007

P. Smolensky. On the Proper Treatment of Connectionism. *Behavioral and Brain Sciences*, 11, 1-74, 1988

R. Spinello and H. Tavani. *Readings in Cyberethics*, 2nd ed. Tandem, 2004

J. Steuer. Defining Virtual Reality: Dimensions Determining Telepresence. *Communication in the Age of Virtual Reality*, 1995

P. Suber. What is Software? *Journal of Speculative Philosophy,* 2(2), 89-119, 1988

L. A. Suchman. Plans and Situated Actions: The Problem of Human-machine Communication. Cambridge University Press, 1987

J. P. Sullins. When is a Robot a Moral Agent? *International Review of Information Ethics,* 6, 24-30, 2006

C. Sunstein. *Republic.com.* Princeton University Press, 2001

C. Sunstein. *Infotopia: How Many Minds Produce Knowledge*. Oxford University Press, 2006

C. Sunstein. Democracy and the Internet. In *Information Technology and Moral Philosophy*, J. van den Hoven and J. Weckert, eds., pp. 93-102. Cambridge University Press, 2008

K. P. Sycara. The Many Faces of Agents. *AI Magazine,* Summer 1998, pp. 11-13, 1998

J. H. Søraker. Global Freedom of Expression within Non-Textual Frameworks. *The Information Society*, 24(1), 40-46, 2008

D. Tapscott and A. D. Williams. *Wikinomics: How Mass Collaboration Changes Everything.* Portfolio, 2008

H. T. Tavani. The Uniqueness Debate in Computer Ethics: What Exactly Is at Issue, and Why Does it Matter? *Ethics and Information Technology*, 4(1), 37-54, 2002

H. T. Tavani. *Ethics and Technology: Ethical Issues in an Age of Information and Communication Technology [2nd ed]*. John Wiley and Sons, 2007

P. Thagard. *Computational Philosophy of Science*. MIT Press,1988.

P. Thagard. Internet epistemology: Contributions of new information technologies to scientific research. In *Designing for science: Implications from everyday, classroom, and professional settings,* K. Crowley, C. Schunn and T. Okada, eds., pp. 465 – 485. Lawrence Erlbaum Associates, 2001

R. Thomason. Logic and Artificial Intelligence. In *The Stanford Encyclopedia of Philosophy (Fall 2003 Edition*), E.N. Zalta, ed., 2003

A. Turing. On computable numbers, with an application to the Entscheidungsproblem, *Proceedings of the London Mathematical Society*, Series 2, 42, 230-265, 1937.

A. Turing. Computing Machinery and Intelligence. *Mind, LIX*, 433-460, 1950

S. Turkle. *Life on the Screen. Identity in the Age of the Internet.* Simon & Schuster, 1995

R. Turner and A. Eden. eds. The Philosophy of Computer Science, special issue of *Minds & Machines*, 17(2), 2007a

R. Turner and A. Eden. The Philosophy of Computer Science. *Minds & Machines*, 17(2), 129-133, 2007b

R. Turner and A. Eden. Towards a programming language ontology. In *Computing, Philosophy, and Cognitive Science*, G. Dodig-Crnkovic, ed. Cambridge Scholars Press, forthcoming a

R. Turner and A. Eden. Philosophy of Computer Science. In *The Stanford Encyclopedia of Philosophy,* E.N. Zalta, ed., forthcoming b

A. Urquhart. Complexity. In *The Blackwell Guide to Philosophy of Computing and Information*, L. Floridi, ed., pp. 18-27, Blackwell, 2004

A. Vedder and R. Wachbroit. Reliability of Information on the Internet: Some Distinctions, *Ethics and Information Technology,* 5 (4), 211-215, 2003

A. Vedder. Responsibilities for Information on the Internet. In *The Handbook of Information and Computer Ethics,* K. Himma and H. Tavani, eds., pp. 339-359. John Wiley and Sons: 2008

P. Virilio. *The Vision Machine.* Indiana University Press, 1994

F. Webster. *Theories of the Information Society*. Routledge, 1995

J. Weckert. Trust in Cyberspace. In *The Impact of the Internet on Our Moral Lives,* R. Cavalier, ed., pp. 95-117. SUNY press, 2005

J. Weckert. ed. *Computer Ethics. International Library of Essays in Public and Professional Ethics Series*. Ashgate, 2007

P. Wegner. Research paradigms in computer science. In *Proceedings of the 2nd international Conference on Software Engineering*, pp. 322-330. IEEE Computer Society Press, 1976

J. Weizenbaum. *Computer Power and Human Reason: From Judgment to Calculation*. Freeman, 1976

J. Whitten, L. Bentley and K. Dittman. *Systems Analysis and Design Methods* [6th ed.]. McGraw-Hill, 2003

N. Wiener. *The Human Use of Human Beings: Cybernetics and Society.* Houghton Mifflin, 1950

L. Winner. Cyberlibertarian myths and the prospects for community. *SIGCAS Computers and Society*, 27(3), 14-19, 1997

T. Winograd and C.F. Flores. *Understanding Computers and Cognition: a New Foundation for Design*. Addison-Wesley, 1987

E. Winsberg: A Tale of Two Methods. *Synthese*, forthcoming

S. Young. *Designer Evolution: A Transhumanist Manifesto.* Prometheus Books, 2005.

P. Zhai. *Get Real. A Philosophical Adventure in Virtual Reality*. Rowman and Littlefield, 1998