Mind the gap : a critique of human/technology analogies in artificial agents discourse

MIND THE GAP

Universitaire Pers Maastricht
ISBN 978-90-5278-795-4
Cover Picture: © Constant, Labyrismen, c/o Pictoright Amsterdam 2008
Cover design: Ilze van Roovert, Piraña grafisch ontwerp

The production of this thesis has been sponsored by:
The Graduate School of Science Technology and Modern Culture (WTMC)
Department of Philosophy, Facultury of Arts and Social Sciences
Maastricht University

MIND THE GAP

A Critique of Human/Technology Analogies in Artificial Agents Discourse

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Maastricht,
op gezag van de Rector Magnificus, Prof. mr. G.P.M.F. Mols
volgens het besluit van het College van Decanen,
in het openbaar te verdedigen
op vrijdag 23 januari 2009 om 14:00 uur

door

Merel Elisabeth Noorman

geboren te Amsterdam op 19 maart 1976

UNIVERSITAIRE PERS MAASTRICHT

Promotor:
Prof. dr. R. De Wilde

Copromotor:
Dr. J. Spruyt

Beoordelingscommissie:
Prof. dr. S. M. E. Wyatt (voorzitter)
Prof. dr. ir. W. E. Bijker
Prof. dr. ir. J. H. Eggen (Technische Universiteit Eindhoven)
Dr. R. P. J. Hendriks
Prof. dr. M. J. van den Hoven (Technische Universiteit Delft)

CONTENTS

ACKNOWLEDGEMENTS

My work in the fields of Artificial intelligence and Science and technology studies has intensified my fascination for the various ways in which we come to understand and shape our world. During the last five years I have been given the opportunity to indulge in this fascination, and explore the questions I had, by studying visions of future human/technology relationships. The image pictured on the cover of this book is part of the *New Babylon* project, in which the artist Constant elaborated one such vision. In this project Constant explored a utopian society of complete creativity and mobility, enabled by the full automation of production. The image, though, is more than a reference to the topic of this book. For me, it resonates with my experience of exploring and wandering through a labyrinth of many perspectives on changing human/technology relationships. I am indebted to the people who have made it possible for me to find my way through this labyrinth.

First of all, I want to thank my supervisors Rein de Wilde and Joke Spruyt for their guidance in my explorations through the myriad of ideas, theories and traditions, and for helping me to take a step back to consider the bigger picture, when I needed to. I am grateful for their feedback, suggestions, critical comments and encouragement during our numerous discussions throughout this project. Whenever it seemed like I had hit a dead-end, my promoter Rein offered me different perspectives, pointed out interesting tensions and made connections that helped me to take the next step. Without my co-promoter Joke this dissertation would have been a lot less readable and I might still be exploring the many paths one can take while thinking about human/technology relationships. She asked the necessary questions and has been a much needed sounding-board.

My research has been about shifting, comparing and contrasting perspectives. To see things from different view points is to notice things that otherwise remain hidden. It has therefore been essential for me to stay connected to the field of Artificial Intelligence. I was fortunate enough to be able to regularly exchange ideas with people that I knew from my years as an undergraduate studying AI, as well as with AI researchers that I have since met. I would like to thank the members of D-CIS lab, and in particular Paul Burghardt, Gregor Pavlin, Martijn Neef and Jan Nunnink, for allowing me to study their project and for taking

the time to explain their research, visions and ideas to me. I am also grateful for the discussions and conversations I had with Ben Kröse, Berry Eggen, Mehdi Dastani, Marten den Uyl, the participants, lecturers and organizers of the 2005 Agent summer school in Annecy and all the other AI researchers who took an interested in this project and offered their points of view.

At the Faculty of Arts and Sciences I found an inspiring environment and welcoming community that enabled me to consider the developments within the field of AI and agent-based computing from an appropriate distance and to situate them in broader social and cultural context. I am particularly grateful for the helpful comments and feedback that I got from the members of the BOTS group, the members of the department of philosophy and my fellow PhD students. My time spent at the faculty has not only been about work though. Saskia (thank you for letting me crash as your place so many times), Alissa, Julia, and Niki made the days at the office a lot more exciting. I was very fortunate to have been able to share the good and the not-so-good times with you. Thanks to Leen, Leentje, Maud, Babbette, Mieneke, Sophie, Ludo, Thijs, Cornelia, Roel, and Patrick for inspiring discussions and the many non-work-related conversations, lunches, drinks and much more. A special thanks to Vivian, Maaike and Martijn; I hope we will continue our dinners and brunches for a long time to come. The revamping of Provum has been a welcome occasional distraction from my project, so thanks to Kees and all the others that participated. I also owe many thanks to Jaqueline, Sabine, Joke, Nicole and Patrick for helping me to sort out the practical issues that come with doing a PhD.

The workshops and summer schools organized by the research school WTMC have been a source of inspiration and great place to meet exchange ideas and thoughts with my fellow PhD students and the guest lecturers working in the field of STS. This book has benefitted from the discussions we had during our days in Ravenstein. I would like to thank the coordinators Paul Wouters, Annemiek Nelis, Sally Wyatt and Els Rommes for all their efforts. I am also grateful for the support that my new colleagues at the Raad voor Maatschappelijke Ontwikkeling in Den Haag gave me during the hectic final stretch of this project.

There are so many people, in the Netherlands and abroad, that have contributed in one way or another to this project. Some of them, however, I owe a special thanks for helping me to write a better book, by reading various parts at different stages of the process: Ruud Hendriks, Jeroen van den Hoven, Mark Coeckelbergh, my new friends Aaron

Martin (who offered to proofread my dissertation even before I met him in real life) and Dorien Zandbergen (I still cannot belief we did not meet earlier given the friends and interests we share) and my old friends Arjen Poutsma (thank you for the dinners and taking so much time out to discuss my chapters), Matthew Smith (our virtual discussions continue to inspire) and Andy Yates (I am happy you're living on this side of the world again). I am very glad I ran into Ilze van Roovert at the last moment. She created the perfect cover for this book.

Finally, this project would not have been possible without the support and encouragement of my family and friends. Their patience, love, and friendship have been invaluable and made it possible for me to continue on this exciting and at times frustrating venture. Thank you mum and dad, Anne Weike, Hernán, Marlén, Evert, Rembrandt and Amber for always being there for me. Linda, Femke, Anna, Mark, Nina, Anna, Mendy, Chantal, Astrid and all my other dear friends thank you for making life fun and keeping my spirits up, even through the harder times. I am very lucky to have these wonderful people in my life.

1. INTRODUCTION: CONNECTING HUMANS AND TECHNOLOGY

Is it possible for a person to love a robot and, vice versa, can a robot love a person? And if so, how would this love be different from the love between two people or between a person and her pet? Will computers one day be able to think or feel? Could these computers substitute for a person? Fed by resonating images of human-like robots and machines with a mind of their own, sketched in novels like Mary Shelley's *Frankenstein* and Isaac Asimov's *Robot Series*, and movies such as *Blade Runner, the Terminator, A.I.* and *the Matrix,* these questions continue to fascinate us. It is hard to avoid thinking about them, as I recently experienced when I was watching the documentary *Mechanical Love*, featuring various androids, *geminoids*, and robotic seals. The documentary explores the idea of emotional relationships between humans and robots. It shows how an elderly lady in a German nursing home becomes very attached to Paro, a robot with the appearance of a cuddly baby seal. Another central character is the roboticist, Hiroshi Ishiguro, famous for his work on life-like robots. One of his most remarkable creations is a *geminoid*, an android built as the roboticist's twin. In an intriguing scene Ishiguro, who is located in another room, speaks to his young daughter through the geminoid. His daughter is noticeably apprehensive about interacting with it; she even refuses to touch the robot. While watching these interactions between people and robots, I caught myself wondering about what separates us from these increasingly life-like electronic devices.

My initial reaction to the documentary illustrates the kind of thinking about intelligent technologies that I want to move away from in this book. More specifically, my objective is to offer a different perspective on the significance of the boundaries between humans and technologies. The possibility of dissolving these boundaries has been a central element of the discourse on intelligent technologies (Franchi & Güzeldere, 2005). Computer scientists, philosophers and sociologists have been concerned with questions such as 'can machines think?'; 'can computers be social?' and 'can robots have emotions?'. These questions have motivated ambitious roboticists, cognitive scientists and other artificial intelligence (AI) researchers to uncover the fundamental mechanisms underlying the human mind (Newell & Simon, 1976; Simon, 1981), to develop robots

with human-like intelligence, social skills or emotions (Breazeal, 2002; Brooks, 2002; Picard, 1997), or to create computational structures that would allow us to upload our minds to these structures (Moravec, 1990). Such ambitions have, in turn, sparked highly charged debates about the possibility of such technologies (Dreyfus, 1992; Franchi & Güzeldere, 2005; Graubard, 1988). Critics have objected to the suggestion that the human mind can be reproduced by machine-like structures. Although interesting in its own way, a preoccupation with the question of whether technologies can be human-like distracts us from a nuanced debate about the possibilities, limitations and risks of current efforts to develop intelligent technologies.

During my years as an undergraduate studying AI, I grew increasingly puzzled by the discrepancy between, on the one hand, the rhetoric of leading visionaries in the AI community and, on the other, the research and development work performed at the time. It was unclear to me how enthusiastic advocates of these intelligent technologies imagined making the leap from clever designs of computational structures to computational entities 'living' beside their human counterparts. I felt that optimistic claims about future intelligent technologies and the debates that followed from these claims had little to do with the majority of the initiatives to develop intelligent technologies. My experiences with AI have led me to consider how visions of intelligent technologies relate to current research and development practices. What is the role of these visions in the design and development of intelligent technologies? What can these visions tell us about future technologies and how they will relate to us?

This book is concerned with visions of future human-like computational entities that serve and work next to humans. More specifically, I will focus on the narratives of AI researchers centered on the metaphor of computers as *intelligent artificial agents*. The term agent has become a common feature in computer science literature, appearing in descriptions of computational structures on various levels, ranging from software architectures, to software development methodologies and languages, to interface design and to more theoretical considerations about computation. In this literature, the term has been used to describe components of software programs and robots as interactive and social entities that perform autonomously in some electronic or physical environment and pursue their own goals.

In drawing analogies between humans and technologies many advocates of artificial agents seem to assume that the gap between humans

and technologies should or will be bridged. This assumption is evident in the recurrent suggestion that the natural next step in technological development is to move computer technologies closer to humans by endowing them with human-like capabilities. AI researchers have projected images of intelligent agents that 'think' for or with a human user, that make decisions, learn, act autonomously, and anticipate and adapt to human behavior. These future artificial agents, so these researchers say, will be capable of supporting or replacing humans in a wide range of activities. They will profoundly change the way we relate to technologies. By learning about us, electronic 'personal assistants' will be able to schedule meetings, find music and books, organize and book trips and manage our e-mail, according to our preferences, habits and interests (Maes, 1994a). Artificial "teammates" will collaborate with their human partners in disaster response efforts and space missions (Sycara & Sukthankar, 2006). Sociable robots will serve as companions or tutors, once they are able to communicate with us in personal and intuitive ways (Breazeal, 2002). In our future homes and offices, we will be surrounded by sensitive and responsive computerized environments regulated by interacting agents (Aarts, Marzano et al., 2003; F. Zambonelli & V. Parunak, 2003).

Although these visions project an enticing future, they raise some important practical and moral concerns. For one, the utopian rhetoric about promises of artificial agents is offset by questions and concerns about the unintended and undesirable effects that the development of these technologies can have. Critics and 'agent researchers' alike have pointed to issues of security, privacy and trust (Luck et al., 2005).[1] Can we trust agents to make decisions concerning our e-mail, our online transactions or the education of our children? Ben Schneiderman, a human-computer interaction (HCI) researcher and a critical observer of agent research, has expressed his concern that the anthropomorphic representations in this field mislead designers and deceive users. "It increases anxiety about computer usages, interferes with predictability, reduces user control, and undermines users' responsibility" (Schneiderman & Maes, 1997, p. 56). One of the most ardent critics, Jason Lanier, even contends that the idea of intelligent agents is "both wrong and evil" (1995, p. 66). It misrepresents what computers can do, but more importantly for Lanier, in order for agents to look smart, people make themselves dumb. He warns against the consequences that

[1] From now on I will use the term agent researcher to refer to individuals advocating and engaging in research on artificial agents.

the development and use of these kinds of technologies can have on our understanding of what it means to be human (see also Hayles, 2005).

In debates about artificial agents and intelligent technologies, my concerns stem from the tendency to lose sight of the metaphorical nature of visions. Concepts like intelligence, agents, and autonomy are more than descriptions. They are metaphorical concepts that mask a host of assumptions and serve a variety of purposes. The promises of and objections to artificial agents underline the normative issues at stake in conceptualizing and developing new technologies. They demand an analysis of the assumptions that underlie the development of artificial agents, and of how these assumptions can affect human/technology relationships. A meaningful discussion about artificial agents therefore requires a further reflection on the metaphors used to conceptualize these technologies and how agents should relate to us.

As I will argue in this book, when it comes to current technological developments, we should be concerned with the conditions under which technologies *should* or *should not* be considered human-like. These questions cannot be answered through an abstract analysis of what computers and humans are and do. It requires a contextualized analysis of human/technology relationships, coupled with a critical interrogation of the metaphors used to conceptualize human/technology relationships. The following chapters will show that such an analysis provides a basis for both a more reflective approach to the development and design of new computer technologies, as well as for a nuanced debate on the social and moral aspects of agent technologies and intelligent technologies in general. In this introductory chapter, I will first discuss the two departure points of this book: the context-dependent nature of human/technology relationships and the role of metaphorical concepts in AI research. In section 1.3 I will address the concept of 'intelligent artificial agents' and 'agent-based computing' as a field of research. The last part of this chapter presents the outline of this book.

1.1 A CONTEXTUALIZED PERSPECTIVE

According to the futurologist and AI researcher Ray Kurzweil, the advent of increasingly intelligent technologies is unavoidable as a result of what he calls the *law of accelerating returns*: "the inherent acceleration of the rate of evolution, with technological evolution as continuation of biological evolution" (p. 7). He foresees an exponential growth of the capacity of information technology as a result of general trends in

current and past technological developments, illustrated by the predictive power of Moore's Law and ongoing advances in the neuro- and cognitive (computational) sciences.[2] As computers are becoming more powerful and capable of performing a wider range of functions they will enable the development of even more advanced technology.[3] In Kurzweil's account technology evolves according to some autonomous inherent logic towards the ultimate end-point of a superior non-human intelligence. Technological evolution will result in increased speed, efficiency, cost-effectiveness and, most of all, 'order'. For Kurzweil human intelligence is merely an imperfect exemplar:

> In fact these future machines will be even more humanlike than humans today. If that seems like a paradoxical statement, consider that much of human thought today is petty and derivative. We marvel at Einstein's ability to conjecture up the theory of general relativity from a thought experiment or Beethoven's ability to imagine symphonies that he could never hear. But these instances of human thought at its best are rare and fleeting [. . .] Our future primarily nonbiological selves will be vastly more intelligent and so will exhibit these finer qualities of human thought to a far greater degree. (2005, p. 378)

Kurzweil's account is on the extreme end of prophecies and predictions about what current efforts to develop intelligent technologies will lead to. However, it illustrates two recurrent and problematic elements in visions of intelligent technologies. First of all, Kurzweil spins a determinist and teleological narrative in which he looks upon technological change as a natural, law-like process that impacts social change (Mackenzie & Wajcman, 1999). Such deterministic and teleological narratives are problematic when it comes to discussing what future intelligent technologies will or should mean and do in relation to humans. They imply that the margins for steering the development of technologies are small. The suggestion that technology developments follow some autonomous logic does not leave much room for a discussion on which technologies we want to develop, and how we want

[2] In 1964 George Moore predicted that the number of transistors placed on a microchip doubled every year. Moore's Law, as it has since become known, predicts that computing power doubles in a fixed period of time (Ceruzzi, 2003). Computer power is measured by the number of transistors that can be placed on a microchip.

[3] In the coming decades, Kurzweil suggests, technological change will rapidly accelerate bringing us closer to the point, that due to Vernon Vinge we now refer to as the *Singularity* (Vinge, 1993). "The Singularity will represent the culmination of the merger of our biological thinking and existence with our technology, resulting in a world that is still human, but that transcends our biological roots" (Kurzweil, 2005, p. 9).

to organize our world. Secondly, Kurzweil's account exemplifies the use of abstract and ambiguous concepts that mask particular assumptions about what it means to be human, what intelligence entails and what kind of technologies are being developed.

A closer look at the developments in intelligent technologies makes the abstract nature of Kurzweil's visions apparent. More than half a century of AI research has resulted in a fragmented but wide-ranging field of research projects with many conceptions of how computational intelligence can be achieved, what it entails or what it should do. Since its early beginnings in the 1950s, the field of AI cultivated the analogies between humans and technologies (McCarthy et al., 1955; McCorduck, 1979).[4] These analogies have inspired research projects aimed at uncovering the processes of the mind. At the same time, most researchers working on intelligent technologies are unconcerned about accomplishing the so-called 'overarching goal' of building an artificial mind. Reflecting on the past fifty years of AI, the chairman of the American Association of Artificial Intelligence, Ronald Brachman notes that "as a whole the field doesn't seem to be making a lot of progress in that direction [true artificial intelligence], even while we make tremendous progress in our specialized areas" (2006, p. 22). Current research projects and problems cannot be captured by a single definition of the topic, theory or methods. Moreover, no single field of research can claim sole ownership of the idea of intelligent computers. Various researchers in a wide range of disciplines have developed their own lines of research and development, in reaction to or building on the theories and goals of the early AI pioneers. These explorations have led to new fields of research including *artificial life* (ALife), *connectionist computing*, HCI and *multi-agent systems*.

An analysis of current technological developments within their historical, cultural and conceptual contexts that acknowledges the interdependencies between humans and technologies will provide a more promising basis for the perspective I wish to develop. I take my inspiration from work in the field of Science and Technologies Studies (STS).[5] In general STS investigates the influence of social, temporal,

[4] McCarthy is generally credited with coining the term Artificial Intelligence. He introduced the term at the Dartmouth conference as a label for a projected new field of science. The conference is therefore often heralded as the birthplace of Artificial Intelligence (McCarthy et al., 1955; McCorduck, 1979).

[5] Different schools of thought are commonly grouped together under the header of constructivist studies of technology and STS. *Actor-network theory* and the *social construction of technology* are two movements that have applied the insights of sociological studies of

cultural, economic or political factors on the development and use of technology, as well as with how technological artifacts shape society (Bijker & Law, 1992; Hackett et al., 2008; Mackenzie & Wajcman, 1999). The literature in this field has highlighted the multidirectional and contingent trajectories of technological developments (Bijker et al., 1987). From an STS point of view, the claim that technological evolution is an autonomous process that impacts society is not only a simplification of the connections between humans and technologies, it also "absolves us from responsibility for the technologies we make and use" (Wyatt, 2008, p. 169).

A growing body of STS literature underscores that technological artifacts are not isolated objects that mean and work the same regardless of why, by whom and in what context they are developed or used. The work of Wiebe Bijker and Trevor Pinch on the *social construction of technology* (SCOT), for example, demonstrates that the *interpretive flexibility* of artifacts leads to different uses as well as to different designs (Bijker et al., 1987). Different 'relevant social groups' have varying criteria for judging what makes a design superior or even workable, depending on, often competing, goals and interests as well as on distinct ideas about what a particular artifact should do. SCOT has drawn attention to the social processes and local effects that shape the design, meaning and use of technological artifacts, which ahistorical, determinist accounts of technologies have left unexplored. Other lines of research in STS have emphasized the need for a simultaneous account of social processes and the efficacy of technological artifacts. In particular, *Actor-network theory* (ANT), as developed by Bruno Latour, Michel Callon, Madeline Akrich and John Law, has insisted on a symmetrical treatment of human and non-human actors in analyzing the relationships between humans and technologies (Latour, 2005; Law & Hassard, 1999).[6]

---

science and knowledge to the study of technology. Although they differ in some fundamental theoretical and empirical commitments, they share two important features. Both originated from a rejection of the traditional linear models of the relationships between science, technology and society. In addition, they both share a commitment to some form of constructivist analysis in the sense that they assume that society and technology are mutually constitutive. For a further historical account of these developments in technology studies as well as more detailed explanations of the different schools of thought see Bijker et al. (1987), Hackett et al. (2008) and Sismondo (2004).

[6] ANT builds on a rigorous application of the *principle of symmetry* in the analytical treatment of technologies, humans and other non-humans. In other words, according to ANT the distinction between humans and technologies is an outcome rather than a given. Law characterizes ANT as a semiotics of materiality that conceives of entities as

From the constructivist perspective provided by STS, technological artifacts are intimately and complexly connected to the elements of their contingent surroundings through processes of *mutual shaping* or *co-construction*. Technologies are not only the causes of social trends, they are also the effects of these trends (Edwards, 1994). To bring these processes into view, STS scholars extend the focus of analysis to *sociotechnical* systems. This term underlines that a technological artifact can only be properly understood in terms of its connections to a larger heterogeneous system that is never merely technical. Modern society, in STS, is conceived of as a *seamless web*, in which "it is never clear a priori and independent of context whether a problem should be treated as technical or as social and whether solutions should be sought in science, economics, or some other domain" (Bijker, 1995, p. 273).

Although various competing theories exist within STS, the shared central tenet that humans and technologies are inextricably linked in sociotechnical systems through processes of mutual shaping provides a departure point for my analysis. It reminds us that research on artificial agents takes place within social contexts, where political, moral and cultural ideas shape the design, use and meaning of these technologies. An awareness of these aspects of technological developments is needed for reflective research and development practices, as has been noted by agent researchers Franco Zambonelli and Michael Luck. In their essay, *Agent Hell: A Scenario of Worst Case Practices*, they sketch a decidedly dystopian image of a future world filled with computational agents (2004). They project a world in which narrowly focused design practices have resulted in a situation where people's lives are at the mercy of incomprehensible, complex, and overloaded computer networks populated by uncontrollable agents. With their story Zambonelli and Luck aim to emphasize the need for rigorous software engineering processes that take into account the social, political and environmental aspects of technological development. The STS point of view provides a basis for such processes.

In addition, a constructivist perspective emphasizes that computer technologies like artificial agents are intimately related to our understanding of what it means to be human. Computers have become such an

---

taking their form and acquiring their attributes in relation to other entities (Law, 1999, p. 4). "In this scheme of things entities have no inherent qualities: essentialist divisions are thrown on to the bonfire of the dualisms" (p.4). Properties of humans or technologies are not to be found or discovered, but they are created in networks of heterogeneous actors.

integral part of human life that it is hard to discuss what it means to be human without making reference to these devices. They have become an essential part of human activity as they extend, support, and form human abilities and link us to the network of people and technology that now defines society. On a conceptual level they have provided an attractive analogy that has profoundly shaped our understanding of human life (Edwards, 1996; Hayles, 1999; Turkle, 2005). Even how we think and speak about future technologies, as Katherine Hayles has argued, shapes our conceptions of what it means to be human. "Whether or not the predicted future occurs as it has been envisioned, the effect is to shape how human being is understood *in the present*" (Hayles, 2005, p. 132). The extent to which computers will be like humans therefore depends on the choices we make in researching, developing and using technologies.

Hence, important questions that need to be addressed are: Why and in what sense do advocates of artificial agents propose that these technologies *should* have human-level intelligence, emotion, morality, agency or some other human-like quality? In this book, I will explore these questions by taking a closer look at the metaphorical concepts that these agent advocates enlist to describe their envisioned technology and its promises. In the following section I will explain my understanding of metaphors and say more on how a focus on these metaphors contributes to my analysis. To this end, I will briefly revisit some of the ideas that originated in the formative years of AI.

1.2 METAPHORS IN CONTEXT

The early decades of AI have come to be associated with the belief that thinking and intelligence are properties of disembodied, symbol-manipulating, information-processing machines. This is in large part the result of the prominence of the *physical symbol system hypothesis* put forward by Herbert Simon and Alan Newell (1976). The hypothesis states that: "a physical symbol system has the necessary and sufficient means for general intelligent action" (p. 161). The theory implies that a digital computer, as a type of physical symbol system, can exhibit human-like intelligent behavior by supplying it with an appropriate symbol-processing program. It also entails that human intelligence can be explained in terms of symbol manipulation.

The physical symbol system hypothesis was to become the heart of early artificial intelligence research, now commonly known as classical AI, or in a slightly more derogative version *Good Old Fashioned AI*

(GOFAI). In its formative years the emerging field of AI was heavily influenced by Simon and Newell's ideas. AI researchers developed a research agenda that prioritized the search for general principles that could serve as a foundation for the development of symbol-manipulation programs.[7] They set out to find formal representations that could accurately describe (aspects of) intelligent behavior in terms of well-defined algorithmic rules applicable to logic-based, symbolic structures (Winograd, 2006).

Critics of AI and the broader field of cognitive science have taken exception to the suggestion that the human mind and computers could be thought of as governed by the same general principles (Graubard, 1988).[8] They have argued against the presupposition that knowledge and intelligence could be captured in computational structures and mathematical or logical models, as the physical symbol system hypothesis suggested. Herbert Dreyfus, one the most vocal critics of the symbol-manipulation approach, objected to the notion of intelligence as disembodied, abstract process (Dreyfus, 1992). According to Dreyfus, AI researchers assumed that human thinking is governed by mechanisms that can be isolated from their context and described in a set of general, objective rules or scientific laws. Drawing on the phenomenological writings of Heidegger and Merleau-Ponty, Dreyfus instead located the necessary conditions for intelligence in the relations between the human body and its environment. Intelligence can therefore not be reduced to context-free principles. In a similar fashion, other critics of AI pointed to a range of proposed inherent properties or abilities that humans have and machines lack, such as emotion, common sense and intentionality.

In his analysis of various objections to AI, Warren Sack argues that critics, including philosophers, anthropologists and disaffected AI practitioners, have tended to recapitulate modernist and humanistic philosophical debates about human nature (Sack, 1997). For example, he points out that critics have rejected the rationalist position of the early AI project by emphasizing the role of the senses, the body and the environment in human cognition. In formulating their critique they have taken an "essentialist stance" of the form: "AI will not succeed because

[7] Simon and Newell were members of an interdisciplinary group of scientists that rallied around the (at that time) new idea that the processes of the mind (rather than the brain, as cyberneticians had proposed earlier) could be generated by computational structures (McCorduck, 1979).

[8] McCorduck gives an account of the heated debates between people that believed it was in principle possible to replicate a mind in a machine and others who were fiercely opposed to the idea (McCorduck, 1979).

humans have but computers do not, and cannot, have one or more of these: bodies, on-going social relationships, neurobiological brains, and, situated, indexical representations of the surrounding environment" (Sack, 1997). As Sack points out, the AI community and related disciplines continue to respond to critiques of 'what computers can't do' by redefining intelligence and appropriating concepts such as social interaction and embodiment as inspiration for new research directions (see also Woolgar, 1987).

However influential the early 'essentialist' debates have proven to be in shaping our understanding of humans and technologies, they have not led to any form of consensus about the nature of humans, human intelligence, cognition or technology, nor are they likely to do so in the near future. Concepts, like intelligence, thinking, and consciousness, in these debates have turned out to be moving targets. If anything, the project of AI has contributed to the diversification of debates centered on the issue of intelligence rather than bringing them to an end. In placing humans and technologies next to each other, as two abstract separate entities in a comparative analysis, participants in essentialist debates continue to redefine concepts as part of their efforts to either draw boundaries or to transgress them. Such a comparative analysis, however, overlooks or bypasses the connections through which humans and technologies co-constitute each other materially, as well as conceptually.

Metaphors

An alternative and more promising way to understand the efforts of AI researchers is to acknowledge the metaphorical character of the concepts used to describe their technologies. The physical symbol system hypothesis was the result of an exploration of the analogy between humans and technologies, as evidenced by Simon's own recollection of the early days:

> When I first began to sense that one could look at a computer as a device for processing information, not just numbers, the metaphor I'd been using, of a mind as something that took premises and ground them up and processed them into conclusions, began to transform itself into a notion that a mind was something which took some program inputs and data and had some processes which operated on the data and produces output. There is quite a direct bridge, in some respects a very simple bridge, between this earlier view of the mind as a logic machine, and the later view of it as a computer.
>
> *(Simon as quoted by McCorduck, 1979, p. 172)*

Simon used the metaphor of the logic machine to think about and explain certain aspects of the mind, and he, in turn, used the mind as metaphor to understand processes of the computer.

The metaphors used by Simon serve as more than ornamental devices or tools of persuasion in rhetoric. Rather, they work on a conceptual level to support a particular understanding of the world and to give shape to this world. The role of conceptual metaphors in the production and generation of knowledge as well as in shaping our actions has been explored more extensively by George Lakoff and Mark Johnson. Metaphors are not just a matter of language, they argue. They have a profound effect on our language, thoughts, experiences and actions (Lakoff & Johnson, 1980). According to Lakoff and Johnson our conceptual system governs our thought and every day functioning. "Our concepts structure what we perceive, how we get around in the world, and how we relate to other people. Our conceptual system thus plays a central role in defining our everyday realities" (p. 3). This system of concepts, they claim, is largely metaphorically structured.[9] Our understanding of a concept is formed through linkages with other concepts that highlight and hide aspects of the phenomenon to which the concept refers. They give the example of the conceptual metaphor *argument is war*. Although arguments and war are different kinds of things, the concept of war *partially* structures the concept of argument. This metaphor allows one to systematically think of and experience the act of arguing in terms of winning and losing, attacking and defending, planning and using strategies. Arguing conceived of as war, however, hides the cooperative aspects of this activity. Similarly, the concept of intelligence provides a metaphor to partially structure the understanding, as well as the experience of computer systems in terms of another more familiar feature of human behavior.

Metaphors have been a topic of discussion and research in computer science (Brooks, 1987; Erickson, 1990). Computer scientists explicitly enlist metaphors to structure their definition of a problem and the

[9] Our normal conceptual system, according to Lakoff and Johnson, is metaphorically structured and grounded in experience and culture. We understand most concepts in terms of other concepts. Some concepts are grounded in the physical experience of the body in interaction with the world. The concept of 'up', for instance, is understood with reference to our motor-perceptual experiences. However, they note that experiences are thoroughly cultural, in the sense that cultural assumptions, values and attitudes are already present in our experiences. "[E]very experience takes place with a background of cultural presuppositions" (p. 57). Experience itself, Lakoff and Johnson note, is partly metaphorical in nature.

envisioned solutions. They have proposed particular metaphors to analyze, design and implement complex software systems, as in the case of *object-oriented programming*, *web services, grid computing* and *multi-agent systems*. The concept of intelligent agents is often explicitly introduced as a new 'design metaphor' for software development (Luck et al., 2004; Wooldridge & Jennings, 1995). In addition, metaphors are used to guide the user in operating the technology (Norman 1999).[10] Concepts that the user is assumed to be more familiar with, such as a desktop or file folders, support a conceptual framework for the design of the interface. The treatment of metaphors in these practices shows little reflection with regards to the context in which they originate. Moreover, the concepts computer scientists use are not always recognized as metaphorical.

A closer look at particular contexts shows that metaphorical concepts acquire particular interpretations that make sense within those contexts. Simon and Newell developed their ideas in an intellectual and scientific climate in which conceptions of human behavior and formal models of information processes, i.e. mathematical or logical representations of these processes, were already linked. They had been studying human problem-solving and decision-making in a setting characterized by a strong belief in the potential of mathematical and logical modeling and simulation techniques in strategic planning (e.g. military strategy) (Edwards, 1995; McCorduck, 1979). Inspired by the interdisciplinary work of cyberneticians, information theorists and game theorists on the study of *communication* and *control*, Simon and Newell used the computer as a tool and metaphor to simulate human problem-solving and decision-making in organizations. According to Simon, it was Newell's work on organizational problems in the Air Force that convinced him that these human activities were fundamentally information-processing activities. During their collaboration they found "a common ground in the study of information processes as a route to understanding human decision making in organizations" as Simon himself puts it (Simon, 1997).

The context in which the notion of a computational mind, as entertained by Simon and Newell, became meaningful can be described as

[10] Donald Norman has argued that the common practice in interface design to develop an interface around a metaphor, such as "the word processor is like a typewriter", is misguided (Norman, 1999). "It is true that a metaphor is appropriate in the initial stages of learning. […] After those first few steps of learning the metaphor is guaranteed to get in the way, because by the very nature of metaphor, the thing being represented by the other isn't the same" (p. 181). He argues that designers should instead make a clear understandable *conceptual model*. I consider such a conceptual model to be metaphorically structured as well, albeit in a less obvious way.

discourse. As a theoretical concept in sociological and philosophical studies of science and technology, discourse provides an analytical tool to study the different understandings of physical and social reality (e.g. Peters, 2006). This conception of discourse refers to more than conversations and debates, or linguistic phenomena. It captures the interdependencies between a set of particular practices and the knowledge that they produce. In his study of the development of the computer and its role in the sociopolitical events of the Cold War period, Paul Edwards uses the notion of discourse to explore the connections between metaphors, technologies and humans. He considers this notion to refer to the "entire field of signifying or meaningful practices [. . .] through which reality is interpreted and constructed for us and with which human knowledge is produced and reproduced" (p.34). Drawing on Foucault and Wittgenstein, he defines a discourse as "a way of knowledge, a background of assumptions and agreements about how reality is to be interpreted and expressed, supported by paradigmatic metaphors, techniques, and technologies and potentially embodied social institutions" (1996, p. 34). A discourse combines heterogeneous elements, such as traditions, regulations, protocols, techniques, languages as well as fragments of other discourses, around an object or objects of knowledge. Concepts, ideas and theories are formed by and in turn form a set of practices. Defined as such, discourses have a certain level of coherence, yet they are dynamically regenerated and changing. Central to Edwards' conception of discourse is the idea that meaning (and truth) is inseparable from interactions between humans and from human interactions with their material environment.[11]

Edwards contends that, like language and social practices, computers as material devices and metaphors are elements of discourse (1996). They shape discourse, but discourse also shapes them. In his historical study of intersecting discourses in the Cold War era, Edwards shows that

[11] Building on Wittgenstein's notion of language games, Edwards conceives of language in terms of actions, rather than representations. Humans acquire the meaning of words and learn to speak a language in interaction with the world and within habitual, instinctual, traditional, and institutionalized patterns of actions. The notion of language games expresses the idea that we learn the meaning of a word through associating it with experiences and actions, or through relating it to other familiar words. We acquire the meaning of the term and concept of 'chair' through a process of repeated activities, where a set of objects are pointed out to us as instances of the concept of chair and where we connect this pattern of activities with the concept of sitting on this object. We learn to employ terms and construct concepts through these language games (Edwards, 1995, p. 34 -37).

emergent fields of science like cybernetics, AI and cognitive psychology were part of a *cyborg discourse* (see also Haraway, 1991). "This discourse is primarily concerned with the psychological and cultural changes in self-imagining brought on by the analogy between computers and minds" (p. 21). The central concepts of computing and information-processing supported new ways of thinking about intelligence, languages and thought, such that they could be applied equally to humans and computers. These new ways of thinking generated a variety of new perspectives, self-interpretations and social roles, which transcended the distinctions between humans, animals and non-living systems. In AI and cognitive psychology the computer was conceived of as mind, while it in turn provided a metaphor for explaining the human mind. Images and ideas about man-machine integration featured as central elements in this discourse. "The word 'cyborg', or cybernetic organism captures the strategic blurring of boundaries inherent in these metaphors. Cyborg discourse, by constructing both human minds and artificial intelligences as information machines, helped to integrate people into complex technological systems" (p. 2). Edward's historical analysis of the broader cultural discourses involved with the development of the computer indicates not only the social construction of technology; it also shows the technological construction of sociopolitical discourse. It demonstrates the role of technology as symbols and as metaphors in the discursive practices that produce and reproduce realities.

Following Edwards, I consider metaphors to be discursive elements that acquire meaning within particular social and material practices, which these metaphors help to shape. My objective, however, is to trace the various, and often conflicting, meanings that metaphors acquire within different discourses. Rather than showing how these discourses are constructed, the emphasis is thus on the use and meaning of metaphors within different discourses. In their study of the dynamics of knowledge, Sabine Maasen and Peter Weingart consider a metaphor in (scientific) discourses to be a "unit of meaning-producing communication" (1995, p. 16). They conceive of metaphors as referring to "the transfer of a concept endowed with a meaning derived from a specific context to another context where it unfolds its transferred meaning" (ibid.). Metaphors are familiar concepts transferred into contexts in which they are unfamiliar. In interaction with discourses metaphors "shift meaning" and produce "new semantics and new pragmatics, new knowledge and new world views even" (Maasen & Weingart, 2000, p.

34). This aspect of metaphors as discursive elements underlies my notion of metaphors.

Metaphors, AI and human/technology relationships

When we consider metaphorical concepts as discursive elements, we see that concepts such as intelligence can become disconnected from their meaning in other discourses and acquire new meanings. The use of the metaphor 'intelligence' in AI research projects does not necessarily entail human-like intelligence. AI researchers Ford and Hayes asserted that "Beginning a textbook on AI with the Turing test (as many still do) seems akin to starting a primer on aeronautical engineering with an explanation that the goal of the field is to make machines that fly so exactly like pigeons that they can even fool other pigeons" (1998, p. 80). They claimed that the famous Turing test serves as an iconographic exemplar of the characterization of the goals of the early decades of AI. Turing's thought experiment, which offered a substitute for the question of whether machines can think, has served in debates about the possibilities of AI as a convenient goal post (Turing, 1950). Ford and Hayes are of the opinion that this misrepresents the ambitions to realize computational intelligence, as it suggests that the single overarching goal is to build an artificial human-like mind.[12] In their view "the scientific aim of AI research is to understand intelligence as computation, and its engineering aim is to build machines that surpass or extend human mental abilities in some useful way" (p. 79). This conception of AI underlines that in order to understand the meaning of concepts used by AI researchers we have to turn towards the discourses in which these concepts are formed.

The use of metaphors in AI and related fields has been previously addressed by members of the AI community as well as by outside observers (Agre, 1997; Hayles, 1999; West & Travis, 1991). These scholars have examined the historical, cultural and theoretical contexts in which AI researchers have cultivated particular metaphors like computational 'minds' and 'rational planning'. The main objective of these scholars has been to challenge influential claims about how computers can be used to explain human nature. Self-proclaimed (former) 'AI person' Philip Agre challenged the metaphors used by early AI research-

[12] Ford and Hayes argue that the aim of AI is to "create a computational science of intelligence itself, whether human, animal or machine" (p.81). They translate this as the "study of how computational systems must be organized in order to behave intelligently".

ers that supported an understanding of cognition as abstract processes in the head (1997). He presented a reconstruction of the ideas, theories and practices of rationalist traditions in AI. His objective was to encourage a more rigorous reflection on the concepts used to understand human nature. The focus in this book, instead, is on the role of analogies between humans and technologies in conceptualizing and designing future computer systems that would enable new kinds of human/technology relationships. By concentrating on visions of changing relationships between humans and technologies, I place less emphasis on research projects concerned with the study of intelligence through computational means.

The connections between humans and technologies have been a central concern within the field of HCI. As Terry Winograd points out, researchers in this field have challenged the suggestion that computers should be more human-like in order to more effectively interact with humans (Winograd, 2006). Instead these researchers have explored more pragmatic approaches to the development of computer technologies. According to Winograd, also a former AI researcher, the emphasis in these approaches is on designing technologies in iterative processes of prototype testing and refinement, rather than on theoretical and "rationalistic" approaches to model cognition in terms of formal symbolic representations (p. 1257). These approaches are concerned with the effects of particular designs on human actions and experiences. Nevertheless, analogies between humans and technologies continue to be a pervasive element of visions of future human/technology relationships and they continue to guide research projects concerned with development of innovative computer technologies. By focusing on analogies between humans and agent technologies, I aim to explore how we can think in a more reflective way about the development of such technologies and their consequences.

For the purpose of my analysis a strict distinction between various fields or between science and engineering would be counterproductive, as it would suggest that the various discourses are unrelated.[13] The interaction between the discourses is highly relevant, as these discourses share a number of metaphorical concepts. The concept *technoscience* provides a more convenient way of thinking about the wide variety of disciplines and initiatives in which analogies between humans and technologies are a central feature, including AI, ALife, HCI and cognitive

[13] In her study of AI researchers, Alison Adam distinguishes between science and engineering to delimit her focus of analysis (1998).

science. Scholars in the field of STS have argued that the distinctions between scientists and engineers are products of the historically grown disciplinary boundaries (Haraway, 1991; Latour, 1987). They instead prefer to speak of technoscience to emphasize that material, techniques, technology are inherently part of the practices of science, just like these practices shape technological development.

To provide a focus for the analysis I concentrate on interrelated concepts centered on the metaphors of artificial agents. Like artificial intelligence, cyborg and adroid, the notion of artificial (intelligent) agents challenges traditional boundaries between humans and technologies. It suggests a reconsideration of the distinction between, on the one hand, humans as entities that initiate actions and, on the other, technological artifacts as passive objects.

1.3 ARTIFICIAL INTELLIGENT AGENTS

In disciplines such as AI, ALife and cognitive science the agent metaphor appears frequently in descriptions of intelligent computer technologies. The ill-defined nature of the concept of agent makes it a convenient container term to support a range of conceptualizations of both humans and technologies. AI pioneer Marvin Minsky, for example, described human intelligence as resulting from the interactions between a 'society' of simple components called agents (1988). Agre, mentioned above, argued that adopting the term *situated, embodied agent* would provide a more promising basis for using computers to explain human action (1997). The term *situated*, as used by Agre, refers to the idea that actions are inextricably linked to physical and social situations. He used the term *embodied* in the phenomenological sense of being and existing as body in the world. In more recent research, the concept of agents has served to describe software architectures, software development methodologies, programming languages and interface design models (Jennings & Wooldridge, 1998). In addition, it has been discussed in more theoretical accounts of computation (Wooldridge, 1999).

The metaphor of computer systems as artificial agents has been extensively explored in *agent-based computing* (Luck et al., 2005). Agent-based computing is a label that acts as placeholder for a wide-variety of approaches to computing, driven by a diverging range of ambitions and goals. It is a comparatively young research area that started showing the outlines of a sub-field within computer science around the mid-to-late nineties of the last century. It was during this period that several now

classic publications appeared, and that the term agent technology found its way into the popular computing press (Nwana & Ndumu, 1999). These early publications presented agents as offering a conceptual framework for such diverging topics as human/computer interaction, distributed large-scale system design and modeling, and simulating complex dynamic systems.

As a result of the growing activity in the field of agent-based computing, the concept of artificial agents is hard to pin down. In one of the first textbooks on agents Gerhard Weiß gives the following definition:

> An agent is a computational entity such as a software program or a robot that can be viewed as perceiving and acting upon its environment and that is autonomous in that its behavior at least partially depends on its own experience. As an intelligent entity, an agent operates flexibly and rationally in a variety of environmental circumstances given its perceptual and effectual equipment. Behavioral flexibility and rationality are achieved by an agent on the basis of key processes such as problem solving, planning, decision-making, and learning.
>
> *(1999, p. 1)*

This rather general definition highlights some recurrent ideas about agents, such as their cellular or atomic nature, their embeddedness in some physical or electronic environment, their flexibility, their autonomy and their reasoning abilities.[14] These qualities however summarize only a small part of the range of features that have been associated with the concept of agents in the literature. Franklin and Graesser attempted to "capture the essence of an agent in a formal definition" by mapping out the various definitions of agents, in order to distinguish them from other computational programs (1997). Agent researchers, they note, have varyingly defined agents in terms of being reactive, autonomous, goal-oriented, pro-active, purposeful, temporally continuous, communicative

[14] In various reviews and surveys we find attempts to define and specify what an artificial agent is. For instance, Luck et al. provide an overview based on their comprehensive study of European agent research programs and commercial and industrial applications. They define an agent as "a computer program capable of flexible and autonomous action in a dynamic environment, usually an environment containing other agents" (2006, p.8). They continue: "In this abstraction, we have encapsulated autonomous and intelligent software entities, called agents, and we have demarcated the society in which they operate, a multi-agent system. Agent-based computing, according to the authors, "concerns the theoretical and practical working through of the details of this simple two-level abstraction" (ibid.).

socially, adaptive, mobile, flexible as well as having a believable personality and emotional state.[15]

Different interpretations of computational agents - originating in different traditions - have given rise to diverging lines of research. One common distinction is that between *personal* or *user agent* and *multi-agent system*. The notion of personal agent offers an enticing metaphor for thinking about the interaction between humans and complex computational systems. Current accounts of research ambitions in this area mirror ideas about helpful computer systems in the tradition of *human-machine symbiosis* or *intelligence augmentation* (Licklider, 1960; Skagestad, 1993). Researchers working in these traditions place the emphasis on developing technologies that enhance human intelligence, rather than on building an isolated computational mind. They set out to develop computer technologies that assist a human user in searching for information on the Internet, scheduling meetings, booking trips and managing her e-mail.[16] Pattie Maes, founder of the Software Agents Group at the Massachusetts Institute of Technology (MIT), has played a key role in popularizing personal agents (Maes, 1994a, 1994b). She described these agents as autonomous interactive software programs acting as an intermediary between the user and the web or the computer. "Agents assists users in a range of different ways: they hide the complexity of difficult tasks, they perform tasks on the user's behalf, they can train or teach the user, they help different users collaborate, and they monitor events and procedures" (1994a, p. 31). Maes envisioned agents that would learn about their users' interests, habits and preferences. Subsequent articulations of this metaphor have emphasized the ability of individual agents to sense and reason about their environment, as well as their ability to interact in a social way with human beings (Dautenhahn, 2002). Research on personal agents has found renewed currency in

[15] Wooldridge and Jennings identified four properties that characterize artificial agents: autonomy, social ability, reactivity and pro-activeness. Thus, agents should be capable of acting without interference while interacting with other agents ("possibly humans"). They should be able to perceive and respond to their environment and "exhibit goal-directed behavior by taking initiative" (Wooldridge & Jennings, 1995).

[16] Maes builds on older discourses that embraced the metaphor of the computer as "digital assistant". She notes that the ideas Alan Kay and Nicolas Negroponte served as inspiration for her visions of software agent. Kay counts as one of the first computer scientists dedicated to the idea of 'personal computers' that children too could work with (Kay, 1972). Inspired by McCarthy's Advice Taker, he put forward an image of intelligent assistants that can "clone their users' goals and then carry them out" (1990, p. 203). Negroponte envisioned the advent of digital butlers in his once best-selling book *Being Digital* (1995).

recent trends like Web 3.0 and the Semantic Web. These trends build on a vision of a next generation World Wide Web that will bring structure to the meaningful content on the Web. This machine-readable content will create "an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users" (Berners-Lee et al., 2001, p. 34).

The second common interpretation of computational agents presents them as interacting entities in a distributed system, referred to as a multi-agent system (MAS). Work on MAS derives in part from *distributed artificial intelligence* (DAI),[17] in which the focus shifted from developing a single problem-solving system to societies of problem-solvers (Weiß, 1999). Agent researcher Katia Sycara defines a MAS as "a loosely coupled network of problem solvers that interact to solve problems that are beyond the individual capabilities or knowledge of each problem solver. These problem solvers, often called *agents*, are autonomous and can be heterogeneous in nature" (1998, p. 80). Although various definitions of MAS highlight different aspects, common elements include modularity, heterogeneity, non-hierarchical distributed organization and (social) interaction leading to emergent behavior. Compared to research on personal agents, the emphasis in MAS research is more on the intelligent behavior of collections of agents, than on the reasoning skills of individual agents. Central concerns are coordination and communication among agents. MAS are generally described as populated by different types of agents with a certain level of autonomy that perform different tasks, such as performing sub-procedures and mediating between other agents (Wooldridge, 2002). In practice, research on personal agents and MAS often overlaps.

Similar to AI research, agent-based computing is driven by the (at times overlapping) 'engineering' objective of building applications and the 'scientific' aim of understanding certain phenomena. For instance, the recent *Agent Roadmap*, drafted as part of the *AgentLink Network of Excellence* that ran from 1998 until 2006, presented agent-based computing as a new paradigm (Luck et al., 2005).[18] The authors of the roadmap identify a range of sub-disciplines of information technologies in which they envision agents to play a central role, such as computer networks,

[17] Other traditions in computing like object-oriented programming and concurrent object-based systems have also heavily influenced current work on MAS (Jennings et al., 1998).

[18] See http://www.agentlink.org (last accessed September 12th 2008) for a description of the Agent Link program.

software engineering, HCI, mobile systems, decision-support systems, information retrieval and management and electronic commerce. They foresee numerous application areas, including medical and health care service, business service, e-science and manufacturing and supply chain integrations. In particular, in areas where distributed sources, owned by different parties and operating on different platforms, are to be integrated, agents are proposed as integration solutions that draw the heterogeneous sources together (Luck et al., 2004; Maes, 1994b).

More explicit analogical links between computational agents and humans are explored in research focusing on simulating intelligence and social organization. For instance, AI researcher and MIT professor Rodney Brooks describes his robots as agents. In his research, he explores complex intelligent behavior as generated through interactions with the environment, rather than by acting through isolated internal reasoning and representations. Dissatisfied with the obsession with high-level planning in traditional AI, he proposed a new development paradigm that places an emphasis on *reactive* behavior realized through simple mechanisms.

> We argue that the *symbol system hypothesis* upon which *classical AI* is based is fundamentally flawed, and as such imposes severe limitations on the fitness of its progeny. Further, we argue that the dogma of the symbol system hypothesis implicitly includes largely unfounded great leaps of faith when called upon to provide a plausible path to the digital equivalent of human level intelligence. It is this chasm to be crossed by these leaps which now impede classical AI research.
>
> *[emphasis in the original] (1990, p. 3)*

Brooks contends instead that intelligence is not a property of an isolated entity that processes internal representations, as is implicit in the physical symbol hypothesis, but is to be found in the behavior of an entity interacting with its environment.[19]

Despite the diffuse and wide-ranging conceptions of artificial agents, a number of concepts are recurrent features in descriptions of these agents. In the following chapters, I will focus in particular on three interrelated metaphorical concepts that are prevalent elements in

[19] Like Brooks, Philip Agre and David Chapman noted that little abstract planning is involved in most human daily activities; hence they deemed it necessary to explore new agent-based approaches as an alternative to the early theories of AI pioneers (Agre & Chapman, 1987). At the same time, for many AI has become almost synonymous with building computational structures as reasoning agents (Alonso, 2002; Sycara, 1998).

definitions of artificial agents: *autonomy*, *(social) interaction*, and *adaptivity*. These three concepts each emphasize a particular aspect of the envisioned agents that distinguished them from other computer technologies, but they also feature prominently in accounts of how these technologies will enable new kinds of human/technology relationships. Through these concepts researchers position the envisioned technologies in relation to humans.

First, the most distinguishing concept used to characterize artificial agents is 'autonomy'. Unlike traditional computer systems, agents are proposed as a new kind of computer entities that are capable of independently operating in dynamic and complex environments, independent of human intervention. Agents should have control over their own actions and should be able to generate and pursue their own goals. Such agents would be able to act on the users behalf without being explicitly instructed how and when to perform particular tasks.

Secondly, intelligent agents are often presented as capable of *(social) interaction* with other agents and humans in a way that goes beyond the rigid input/output strategies of conventional computers. In order to autonomously perform tasks agents should be able to interact with the environment and other agents. The interaction metaphor is a relatively recent addition to AI, as it emphasizes the connections between computers and their environments, as well as between computers and humans (Agre, 1997). Human-to-human interaction is one particular interpretation of the interaction metaphor that agent researchers have used to conceptualize human/agent relationships. It underlies many of the definitions of personal agents.

Finally, various definitions of agents characterize these technologies as more *adaptive* than conventional computer technologies. In order to autonomously operate in dynamic environments agents or collections of agents do not only have to interact with their environment; they have to adapt to unanticipated events within this environment as well. Unlike conventional computers, the envisioned adaptive agents would not require a full specification of the tasks they have to perform. Rather they learn or reconfigure their internal structure in response to the contingencies of their environment. For human/technology relationships this entails that more adaptive computational agents would be able to adjust to the activities, habits and preferences of their human users.

I will analyze the three concepts within the contexts in which they acquire meaning. In the following chapters I consider the construction, functions and changing interpretations of these metaphorical concepts in

the literature on agent technologies. The focus is, in particular, on how these concepts constitute representations of human/technology relationships, on the consequences of these representations, and on the conflicts between particular interpretations. Through this analysis, I aim to bring back into focus questions and choices presented by the development of these technologies that have so far been largely overlooked.

1.4 OUTLINE OF THE BOOK

This book continues in the direction set out by sociological and anthropological studies of intelligent technologies. Sociologists and anthropologists in STS have shown that a closer look at the rhetoric and practices of AI researchers provides grounds for further discussion on how intelligent technologies are being developed. The anthropologist Lucy Suchman, for example, has dealt with the question of "what understandings of the human, and more particularly of human action are realized in initiatives in the fields of artificial intelligence and robotics" (2008, p. 144). In her book *Plans and Situated Action,* she addressed the problem of human-machine communication (1987). On the basis of conceptual analyses and ethnographic studies of a group of AI researchers at Xerox PARC, she argued that these researchers worked from particular reductionist conceptions of human action and planning. They conceived of actions as governed by plans that result from rational disembodied problem-solving processes. This conception overlooks the ways in which human activity is *situated* in social and material environments. Human actions and knowledge, according to Suchman, are continuously formed in interaction with a complex world of objects, artifacts and others actors. Context is therefore crucial to the understanding of action. Suchman's analysis resonates with other STS literature on AI. A shared central theme in this literature is the critique of the modernist conception of the human as autonomous, rational individual, in which 'the body', 'the social' and 'the cultural' are systematically erased from notions of knowledge, cognition and action (Adam, 1998; Collins, 1990; Forsyth, 1993).

Studies of AI in STS have made important contributions to the main philosophical debates about the possibility of artificial intelligence. In particular, they have addressed some issues that have been underexposed. Through their focus on the social aspects of these technologies they have drawn attention to the relation of the products of AI research

to humans and how this relation is represented in AI initiatives. They have studied how the particular conceptions of humans built into technologies affect how humans interact with these technologies (Suchman, 1987), the role of AI technologies in social processes (Collins & Kusch, 1998; Edwards, 1994), as well as their influence on how we think about what it means to be human (Edwards, 1995; Hayles, 2005).

Like the earlier STS studies of AI, this book is concerned with how and why researchers of intelligent technologies construct visions of intelligent technologies, as well as with the consequences of these visions. The focus, however, is on an exploration of the heuristic role of the different and often conflicting visions of new kinds of human/technology relationships in the development of agent-based technologies. Although a critique of the representation of knowledge, cognition, actions and other features traditionally associated with humans is part of this investigation, it serves my objective to trace the construction, functions and changing interpretations of metaphorical concepts as they appear in different discourses. An investigation of the heuristic role of visions in agent research allows me to address the following questions. Why does the analogy between humans and technologies continue to inspire research in this area, despite the recurring critiques of efforts to build intelligent technologies (Chapter 2). How do visions of converging humans and technologies work in different contexts (Chapter 3)? How do different interpretations of the metaphors that constitute these visions interact and where do conflicts emerge (Chapter 4)?

The particular conceptions of humans, as well as of human/technology relationships, become visible when placed against a background of other theories of human-technology relationships. I draw on literature from various disciplines, including STS, the philosophy of technology, HCI, and cognitive science, to explore different perspectives on how humans and technologies become connected. These perspectives allow me to consider metaphorical concepts within some of the *problem domains* that agent researchers have been concerned with. The term problem domain in AI research is often used as shorthand to delineate the topics of research projects (Rich & Knight, 1991). A problem domain refers to the focus on a certain field of interest (e.g. theorem proving, medical diagnosis, or speech recognition) in which problems are identified that an AI system should be able to solve. I use the term in a slightly modified sense to highlight the problems identified by agent

researchers in current relations between humans and technologies that the envisioned artificial agents are supposed to solve.

In the following chapters, three problem domains provide a setting for my analysis of the use of metaphors in agent research. The first problem domain concerns the interaction between humans and technologies at the interface (Chapter 2). Agent researchers have suggested that socially interactive agents will provide a more 'natural' and 'intuitive' interface (Breazeal, 2002; Lieberman & Selker, 2000; Maes, 1994a). A key assumption in this type of vision is that in order to improve the human/technology relationship the interactive competences of technological devices need to be leveled with those of humans. The second problem domain centers on the relations between humans and technologies on a cognitive level (Chapter 3). One of the primary motivations driving agent research is to develop intelligent agents that would be able to independently operate in and adapt to unknown, complex and unpredictable environments. Adaptive agents with enhanced cognitive abilities, their advocates claim, would be able to take over increasingly more cognitive tasks, and would extend human cognitive abilities. Finally, issues concerning the delegation and distribution of control, responsibility and accountability between humans and intelligent technologies are the focus of the third problem domain (Chapter 4). The prospect of increasingly complex and autonomous technologies has generated concerns about the loss of control. What happens when things go wrong? Can we still hold humans responsible for accidents caused by an opaque incomprehensible computer system? The concept of autonomy has been a topic of discussions in agent-oriented research projects that aim to address these problems. Some agent researchers have argued that the solution lies in the development of autonomous moral agents (Allan et al., 2000).

The three problem domains are interrelated and cannot be strictly separated. Enhancing the cognitive competences of technologies, for instance, is often a central element in modeling socially interactive agents. However, within the three described domains agent researchers have explored the analogy between humans and technologies from different perspectives, highlighting particular aspects of human behavior and constructing varying conceptions of how humans and technologies should relate. In particular, researchers have enlisted the three earlier mentioned metaphorical concepts - autonomy, social interaction and adaptivity - to shape their conceptions of the kind of agents that would offer a solution to problems in these domains. Agent researchers have

also focused on concepts such as (physical) embodiment and affect to consider how human/technology relationships can be 'improved'. These concepts, however, are less common features of agent definitions.

The focus on the metaphorical character of concepts requires a careful maneuvering in this book, as my aims are analytical, not foundationalist. I draw on a number of theories and ideas from a range of disciplines that have challenged common sense or traditionally rooted meanings of particular concepts. Hence, it is a tricky balancing act to keep the different interpretations of these concepts clearly separated. Nevertheless, I hope this book will show that it is worth the occasional effort in order to re-evaluate the terms of the debate about the possibilities, limitations, and risks of the development of intelligent technologies.[20]

The next chapter takes a first step in the analysis by taking a closer look at the rhetorical aspects of visions of agent technologies concerned with the interaction problem domain. In particular, it explores the persuasive force of these visions by analyzing what they represent and what they hide. The question that this chapter tries to answer is: What makes visions of leveling humans and technologies so pervasive and persistent? The central focus is on visions that recast the human/agent relationship into one between two intentional agents, the computational agent being 'more than a tool'.

In Chapter 3, I turn my attention to visions that figure humans and technologies as similar entities in symbiotic systems. This chapter looks at the instrumental role of metaphors in research and development practices and highlights the different meanings that metaphors acquire in different contexts. The concept of adaptivity and self-organization provide a central focus in this chapter. I will analyze these concepts in the context of an example of a project that combines the elaboration of a vision with the development of a prototype (Storms, 2004b). The cognitive science theory of Distributed Cognition introduced by Edwin Hutchins (1995) provides an alternative interpretation of the processes that bind human beings and technology together in cognitive activities. The objective of this analysis is to examine some of the problematic aspects of taking concepts 'out of context'.

[20] One tricky aspect of discussing metaphorical concepts is the use of quotation marks. I will use single quotation marks to emphasize the contested nature of a concept. However, to ensure readability, I will not use them when it is clear from the context that a concept has multiple interpretations. Thus, I assume that terms such as intelligence, emotion, agents, problem-solving, autonomy etc. are concepts that have varying meanings in different contexts.

Issues concerning the distribution of control, responsibility and accountability between humans and agent technologies set the stage for the discussion in Chapter 4. Based on an analysis of two different meanings of autonomy that are confronted in discourses of autonomous agents relating to humans, this chapter considers the conflicts that different interpretations of metaphorical concepts can generate. I consider to what extent it matters how we speak and think about future technologies.

In the fifth and final chapter of this book I will draw the discussions together and reflect on the implications of my approach with regards to debates about the possibilities, limitations and risks of intelligent technologies as well as to research practices.

2. MORE THAN TOOLS?

At the robot/art STRP festival 2006 in Eindhoven in the Netherlands[1] I encountered an installation called Spatial Sounds.[2] This installation was composed of a spinning mechanical arm with a loudspeaker attached at the end. The arm sensed the presence of people in the room and their distance from it. It made a varying pulsating noise, ranging from a subdued purring to a very loud crackling noise. The arm engaged in a game of attraction and repulsion with its public. Sometimes it would slowly approach a person standing nearby, while softly purring, only to suddenly retreat and spin violently, making a dreadfully loud noise, seemingly repelled by the person. Standing there, watching this relatively rudimentary technical construction and its response to me, I was surprised by the attraction of this spinning arm and its invitation for interaction. I felt drawn to it when it approached me, in what seemed a curious manner, responding to my presence by varying its noise production. But what struck me most was the eerie sense of rejection I felt when the arm fiercely spun away from me, in what appeared to be disgust or annoyance. Its basic feedback loops and limited means of expressions (variation in noise and spinning) were capable of eliciting an odd sense of *otherness* in me. In its interactions with me it created an illusion of intentions or mental and emotional states, such as curiosity and disgust, as if it were alive.

The Spatial Sounds installation shows that how humans relate to a device or artifact is determined by more than their instrumental value; they are more than functional extensions of humans. This becomes particularly apparent when we look at the connections between humans and technology at the *interface* level. The interface between a human and a technological device is the point where they directly affect each other. This is the place where information is transferred and transformed between the two entities. The various ways in which this connection takes shape characterize the different relationships between humans and technology. The relationship between me and the Spatial Sound system is different from the one between me and a tool like a hammer, or between me and the laptop on which I am writing this chapter. They differ both in how I experience them as well as in how they function in

[1] http://www.strp.nl/

[2] http://www.evdh.net/index.html

relation to me. At the interface level, technology takes a particular position or role in relation to humans. One way of describing the role of computers is in terms of the tool metaphor, by characterizing the computer as a device or artifact serving as a functional extension of human cognitive action (Edwards, 1996).

The last few decades have seen a steady growth of research focused on finding new ways for human users to relate to computers and technology, with the aim of 'improving' the interaction of humans with computer devices (Hoffman et al., 2002). This particular niche of computing research explicitly includes the connections between humans and technologies as key element of development methodologies for computer systems.[3] Unlike the so-called *technology-driven* approaches that have traditionally been more prevalent in development and design, a central concern in this *human-centered* research is improving the 'fit' between technologies and the natural inclinations of humans.

The concept of artificial agents has found its way into the niche of human-centered research. In particular, it provides a conceptual basis for visions that suggest the development of computer technologies that will be 'more than a tool' to be operated or used. These visions project a future in which agents become our 'collaborators', 'electronic butlers' or 'invisible intelligent assistants'. Agents would be electronic entities capable of 'thinking with' the human, rather than passive objects to be brought to life by human action. A key assumption is that in order to improve human/technology relationships the interactive competences of technological devices need to be leveled with those of humans. To give shape to these conceptions of human/technology relationships, agent advocates frequently draw the analogy with human *social interaction* or *communication*. Agents are conceived of as entities that anticipate, and adapt to human activity, such that they become seamless, invisible extensions of their human users. Alternatively, they are presented as human-like social and animate entities.

Rhetoric about agents as 'more than tools' mirrors the recurring theme in computer science that presents the leveling of humans and technologies as a natural and logical trajectory of technological development. It echoes familiar debates about promises of human-like robots

[3] Various fields of research and sub-disciplines can be included in this niche, such as HCI, *human-centered design*, *computer-support collaborative work* and *cognitive engineering*. Hoffman et al. provide a taxonomy of the what they call "approaches to the design of complex sociotechnical systems", i.e. those approaches that explore various human/technology configurations with the emphasis on extending and amplifying human cognitive, perceptual and collaborative capabilities (Hoffman et al., 2002).

and machines that relieve humans from their burdening tasks and that extend the space of possibilities for humans. Visions of electronic 'intelligent assistants' or 'team members', therefore, offer a focal point to deepen our understanding of the persisting idea of leveling humans and technologies. In particular, they provide a context in which to address the question of why this is such a persuasive, but problematic theme.

In this chapter I look at the rhetorical strategies that agent advocates enlist to build their visions of improved human/technology relationships around metaphorical concepts like social interaction and communication. I examine how these strategies give shape to these visions by analyzing what these visions represent and hide. In particular, I will concentrate on teasing out the assumptions that support these visions. What, for instance, constitutes 'improvement'? Why would an electronic butler or quasi-human computational entity be better than a tool? What role do humans play in these visions? An analysis of the assumptions shows that advocates of socially interactive agents adopt a narrow decontextualized view of possible human/technology configurations. The final part of the chapter presents more contextualized relational perspectives that reveal a richer landscape of possible human/technology configurations. A preoccupation with blurring boundaries leaves little room for exploring these configurations as alternatives. It therefore leads to problematic design models and distracts from the questions and choices that new technologies pose.

2.1 COMPUTATIONAL BUTLERS AND TEAMMATES

> The new interface paradigm brings us closer to Olimpia's glassy stare: instead of space, those zeros and ones are organized into something closer to an individual with a temperament, a physical appearance, an aptitude for learning – the computer as personality, not space. We call these new creatures – these digital 'personalities' – agents.
>
> (Johnson, 1997, p. 176)

In computing, the term human/computer interaction refers to various styles of *interfacing* with computers. For many conventional systems, this means that the human user provides specific inputs to the computer through its interface (e.g. text-based command line, buttons, pull-down menus, windows, etc.) in order to obtain a required result from the machine. The computer is used as a tool, as a passive object that produces an output based on well defined operations and in response to

a user's input.[4] It is up to the human user to enable and maintain the interaction. *Human-centered computing* focuses on exploring ways in which this interaction can be improved (Hoffman et al., 2002).

The idea of human-centered computing is almost as old as the digital computer itself and has formed a central theme in a wide variety of research approaches.[5] It has inspired research projects that have explored the functional as well as experiential aspects of human/technology interactions. Some research in this niche concentrates on developing models and methods to design "user-interfaces" that enhance the *experience* of the interaction, on a cognitive, physical as well as emotional level (Norman, 1999; Weiser, 1991). Such efforts aim to make interfaces 'attractive', 'intuitive' or 'easy to use'. Other research projects look beyond the interface for ways to improve interaction. Human-centered research projects have, for instance, (empirically) studied how humans (collectively) do things with technologies in order to guide the search for optimal 'task-allocation' between humans and technologies (Hollnagel, 2003).

Human-centered computing does not necessarily entail the development of technological devices that imitate humans.[6] However, the call for the development of artificial agents as an approach to human-centered computing tends to be based on the assumption that computer devices should take a more active role in establishing interaction with humans. The notion of computer systems that perform their tasks without having to be instructed explicitly and unequivocally in formal languages appeals to the imagination. The concept of artificial intelligent agents lends itself particularly well to the conceptual exploration of this idea. It suggests an added interactive quality about these machines, making them more 'alive' and capable of speaking for themselves (Erickson, 1997). Rather than being passive objects to be brought to life by human action, the notion of agents allows for the conceptualization

[4] The extent to which technologies are indeed passive tools is debatable, as will become clear later on (Verbeek, 2000).

[5] Although human centered computing is often presented as a new emerging paradigm, the ideas behind it originate in the early days of computing. For instance, Simon and Newell at the RAND cooperation initially worked on methods to improve the way human operatorworked with technology (McCorduck, 1979). Other early visions of improving human computer interaction included Joseph Licklider's "man-machine symbiosis" (1960) and Engelbart's "augmentation of human intellect" (1963).

[6] The field of Computer Supported Work Collaborative specifically addresses the question "how collaborative activities and their coordination can be supported by means of computer systems"(Carstensen & Schmidt, 2003).

of computational systems as capable of taking a pro-active role in interactions with humans. Interaction, in such scenarios, becomes a more reciprocal activity between the human and the computer.

One of the most widely-discussed applications of the agent metaphor for interface design is the vision of *personal agents* (sometimes referred to as *interface agents* or *user agents*) as elaborated by the MIT researcher Pattie Maes. Maes advocated and popularized the notion of computational agents as *personal assistants* or *butlers* that help or cooperate with the user to operate the complex systems lying behind the interface.

> Instead of user-initiated interaction via commands and/or direct manipulation, the user is engaged in a cooperative process in which human and computer agents both initiate communication, monitor events and perform tasks. The metaphor used is that of *personal assistant* who is *collaborating with the user* in the same work environment.
>
> (Maes, 1994a, p. 31)

Maes presents agents as autonomous interactive software programs acting as an intermediary between the user and the computer or information networks, which perform tasks *on the user's behalf*. Like real butlers, she contends, assisting agents are most effective if they know their master's preferences without asking. Maes' enticing and optimistic account of interface agents conveys a rhetoric that is still wide spread in agent discourse today. This rhetoric presents artificial agents with enhanced social knowledge and communicative skills as the 'natural' and 'necessary' next step to 'improve' human/computer interaction.

To characterize how artificial agents can improve human/computer interactions, researchers have highlighted different aspects of human communicative and social interactions, such as personification, verbal and nonverbal speech acts, shared understanding, and anticipating behavior. They have argued that humans and computer technologies can and should engage in interaction based on shared understanding in order to coordinate actions and achieve mutual goals. These varying emphases generate different and sometimes conflicting conceptualizations of how humans and technologies should interact. On the one hand, agents are described as entities to negotiate and converse with. On the other, researchers characterize agents as assistants or delegates that serve as invisible extension of their human user. To illustrate the variety in agent conceptualizations, I will briefly discuss three approaches to the development of agents that build on the metaphor of social interaction.

Personified agents

One rather literal interpretation of bringing computers closer to humans is exemplified by the ambition of some agent enthusiasts to build computers that exploit the anthropomorphic tendencies of humans in their interactions with computers. This idea is based on a persistent conviction that a more 'natural' way of interacting with technology is facilitated by appealing to the human inclination to attribute human qualities to animals, machines and artifacts (Ford & Hayes, 1998). Throughout the centuries a range of human-like mechanical devices have been developed in the form of *automata* or *androids* that could for instance play music, write or serve tea (Wood, 2002). These devices were made to "arouse interest through their visual appeal and then to inspire surprise and awe through the apparent magic of their seemingly spontaneous movements" (Encyclopædia Britannica, 2007).[7]

More recently, the idea of the *personification of agents*, by giving them a 'face' and a 'personality', has been the topic of several research projects. The most captivating and popular example of this is the recent surge in research into building 'believable' *humanoid* robots, i.e. robots with an appearance based on the human body or some aspects of it. Asimo and Qrio - the astronaut-resembling walking, talking, dancing and singing robots - are examples of such robots.[8] A recent addition to the growing collection of humanoid robots is the *geminoid*. Hiroshi Ishiguro's research group at the Japanese ATR Intelligent Robotics and Communication Laboratories, has developed this "real person-based android", or remotely controlled robotic "twin" of Ishiguro, to study the phenomenon of experiencing "human presence" (Ishiguro & Nishio, 2007).

Personified agents do not have to be robots. They also appear as graphical user interfaces. The infamous Microsoft Bob and its successor Clippy - the pro-active help assistant of Microsoft Office- are two examples of (unsuccessful) attempts to use animated graphical user interfaces to improve human/computer interaction. More sophisticated personified agents appear in the field of *intelligent tutoring systems* (ITS). ITS are systems that are intended to automate a larger part of the tasks normally performed by a human tutor. Several research groups endow

[7] Well known examples include the 18th and 19th century automata Japanese tea-serving mechanical puppets. These were a type of mechanical devices called Karakuri ningyo, which were built for various religious and entertainment purposes. Another example is the Scribe: a mechanical doll, developed by Pierre Jaquet-Droz in 1722, that could write.

[8] For Asimo see http://world.honda.com/ASIMO/ and for Qrio http://products.sony.co.uk/sony_qrio.asp.

these software programs with *avatars* that speak and exhibit facial expressions (Graesser, VanLehn et al., 2001; Koedinger et al., 1997). These different approaches to building personified agents are based on the assumption that humans have a natural tendency to anthropomorphize artifacts and animals. The idea behind such approaches is that developing computer systems that exhibit 'social cues' would be easier to operate, as they would automatically trigger the desired response in humans (Reeves & Nass, 1996).

Electronic teammates

Other agent researchers take the simulation of human interaction one step further. These researchers insist that artificial agents should not only trigger social and emotional responses, but the agent should also be capable of responding appropriately to this behavior. The emphasis in their research projects is on the simulation and formalization of social skills to achieve a more reciprocal interaction. One suggested approach is to model human/technology 'collaboration' on human-style communication. A recent trend is to positions humans and robots or computers systems as operating and 'collaborating' in *teams* (Breazeal et al., 2004; Nourbakhsh et al., 2005; Sierhuis et al., 2003). This trend is apparent in research domains such as space exploration, warfare, and disaster response (Bradshaw et al., 2003; Christofferson & Woods, 2002). The notion of 'human/agent teams' has also been suggested as a useful metaphor in applications areas like training in virtual environments, and personal information management (Scerri et al., 2002; Traum et al., 2003). This model of human/technology interaction receives increasingly more attention in space and military research projects as well as in research concerned with disaster response support with the aim of integrating human and technological activities in new ways.[9]

Some of the research done on 'social robots' by the Robotic Life research group at the MIT Medialab run by Cynthia Breazeal, provides an illustrative example of projects that embrace the communication metaphor to give shape to human/technology collaboration. In her current project, the Leonardo Robot, Breazeal and her colleagues explore the idea of teaching a robot through communicative actions, modeled on formalized descriptions of human communication (Breazeal et al., 2004;

[9] The NASA Ames Research Center has launched a project to experiment with building robots modeled on the idea of team members. One member of the research project was recently quoted in the New Scientist as saying "The big question is whether we should make a better tool or a teammate. […] It's a very different kind of relationship" (Biever, 2007).

Breazeal et al., 2005). Breazeal, a former PhD student of Rodney Brooks, envisions communication between humans and robots as a process of entities continuously updating their internal model of the other entity. This process is based on verbal and non-verbal gestures, unconstrained by the limitations of the conventional rigid input/output mechanisms of robots and computers. She characterizes learning as a team effort in which robots need to be able to understand our intentions, beliefs, desires, and goals, and must communicate their own "set of intents and goals to establish and maintain a set of shared beliefs and to coordinate their actions to execute the shared plan" (Breazeal et al., 2004, p. 316). The notion of a robot as tool, then, is no longer sufficient for Breazeal and her colleagues. Instead these robots become separate entities that no longer fit a "master-slave arrangement" in which the robot is an instrument that humans operate (p. 320). The research team envisages a partnership in which robots "work jointly" with humans "as in the case of collaboration" (ibid.).

Invisible electronic assistants

In the conceptions of human/agent relationships I have discussed so far, agents are represented as human-like entities that simulate some *observable* aspect of human social interaction, e.g. natural language, facial expressions, or maintaining a shared understanding through verbal and non-verbal communication. Some agent researchers give a slightly different interpretation to the metaphor of social interaction to describe envisioned human/technology relationships. They put less emphasis on observable social behavior. Instead, they portray agents as sensitive and responsive entities that can anticipate, understand and adapt to human goals and intentions, and as a result provide a 'seamless' interface. This conception of artificial agents frequently appears in the recently formulated visions of future information societies that go under the header of labels such as *ambient intelligence*, *ubiquitous computing*, *pervasive computing*, and the *disappearing computer* (Aarts, Collier et al., 2003; ISTAG, 2003; Kostakos et al., 2005; Weiser, 1991). These visions foresee a society in which computers are no longer isolated, clumsy, big machines, but integrated and pervasive 'invisible' elements of the environment. Consumer appliances, office automation, health care technologies and other electronic devices and appliances will "fade into the background" and access to information, communication and knowledge will become ubiquitous (Weiser, 1991, p. 94). One unifying feature is the notion that the increasing pervasiveness and connectedness of electronic computing

demands new design paradigms enabling a human/technology interactive experience that is 'intimate' and 'intuitive'.

The idea of artificial intelligent agents as a promising paradigm for human-centered computing is particularly prevalent in ambient intelligence (AmI), as presented and elaborated by the European funded Information Societies Technology Advisory Group (ISTAG), and institutions like the Dutch Philips research lab and the German-based Fraunhofer Institute. AmI builds on the concept of ubiquitous computing (UC), originally proposed by Mark Weiser. UC promotes the idea of developing technologies that are unobtrusive augmentations of the existing environment (Ducatel et al., 2001; Weiser, 1991, 1993).[10] AmI, as presented by ISTAG, embraces this attitude, but presents 'intelligent intuitive interfaces' as a third pillar in addition to ubiquitous computing and communication:

> The concept of Ambient Intelligence (AmI) provides a vision of the Information Society where the emphasis is on greater user-friendliness, more efficient services support, user-empowerment, and support for human interactions. People are surrounded by intelligent intuitive interfaces that are embedded in all kinds of objects and an environment that is capable of recognizing and responding to the presence of different individuals in a seamless, unobtrusive and often invisible way.
>
> *(Ducatel et al., 2001, p. 1)*

Intelligent, intuitive interfaces include technologies that can be operated through speech and gestures. In addition, AmI stresses the need for *context-sensitive* and *personalized interfaces* that are adaptive to individual users. Interfaces should be able to recognize, know and sense human beings, their environment, and other devices (Punie, 2003). Such interfaces would provide a more natural form of interaction that does "not involve a steep learning curve" and that is "relaxing and enjoyable" (Ducatel et al., 2001, p. 11).

The ISTAG presents one scenario in which a communication device "Digital-me" serves as a personnel assistant in handling all technologically mediated communications (Ducatel et al., 2001). The context-sensitive communication device learns the context-dependent preferences of its human user, such that it can decide when a call from a human or another communication device should be passed through to

[10] Context-sensitive technologies are a central feature in Weiser's vision of Ubiquitous Computing. However, he maintains that the emergence of ubiquitous computing does not depend on breakthroughs in AI. Applications do not need to be intelligent to utilize information about their location more effectively (Weiser 1991).

the user, as well as when and how it should take care of the call itself (Ducatel et al., 2001). Such a scenario echoes the more widely advocated idea of intelligent agents that work in the background on our behalf.

Like the notions of personified agents and agents as teammates, the idea of unobtrusive electronic assistants suggests that the social and communicative competences of technologies should be leveled with those of humans. Henry Lieberman and Ted Selker, two vocal proponents of using the concept of intelligent agents for the design of 'unobtrusive' interfaces, provide an illustration of what such a leveling should entail. They propose to shift "some of the burden of dealing with context from the human user to a software agent", by making them sensitive to such things as time, place and the preferences or skills of the user (2000, p. 620).

Problematic visions

Appealing as the agents projected by the three discussed approaches might seem, the concepts of communication and social interaction remain abstract metaphorical concepts that hide a number of contentious assumptions. One assumption underlying the call for responsive and anticipating technologies is that these technologies will empower their human users, as they provide transparency and reduce complexity. However, visions of socially interactive agents leave unexplored the additional demands imposed on the human user. In order for a human user to work with an application that anticipates her behavior in the efficient way the designers intended, she is in fact actively limited in her range of actions. If the agent performs actions on the user's behalf without the user's explicitly instructions, the user will have to accept the abstract representation of herself. In other words, if I allow a news filtering agent to find interesting news for me, I will have to accept that the news it finds for me fits within a profile and everything that falls outside of this profile I will not get to see. Thus, although it might save me time, it limits my flexibility and options.

Furthermore, leveling agent competences with those of humans can lead to new time consuming and unfamiliar actions and place new demands on the human user's skills of appropriation. This becomes clear when we consider what it would take, for instance, to "work with" Leonardo as Breazeal intended. To enable the robot to perform human-like communication acts, Leonardo comes equipped with cameras, computer vision algorithms, speech understanding software modules, an attentional system, "collaborative task-oriented conversation and gestural policies", and more (Breazeal et al., 2004). Nevertheless, Leonardo is

only capable of a very limited range of conversational actions. To implement conversational policies, formal models are defined based on philosophical and socio-psychological theories of human behavior.[11] The robot 'learns' from the human how to perform tasks (e.g. pushing a button that is placed in front of it) to the extent that the formalizations can model gestures and utterances. The robot in turn conveys its internal state to the human tutor by performing actions, and through a limited set of facial and non-facial gestures. It can, thus, only understand human language and gestures if it fits into a specific pattern. In order to converse in a 'natural intuitive' way with the robot, the human has to go through a process of learning to speak the 'right' language for Leonardo to operate correctly, for it cannot make sense of natural speech with all its ambiguity and hidden meanings. A disproportionate part of establishing a smooth interaction, therefore, remains the responsibility of the human.[12]

Objections to the decontextualized and reductionist conceptions of humans, as exemplified by visions of technologies as 'more-than-tools', are repeatedly put forward. They reappear in new guises in response to new approaches and motivations for making technology more human-like. Suchman criticized the use of the social interaction metaphor by AI researchers in the late 1980s. Her critique was based on a study of the efforts of her colleagues at Xerox Palo Alto Research Center (PARC) to develop an intelligent interactive interface to a copying machine (Suchman, 1987). She pointed to the tendency of researchers working in the cognitive sciences to erase the human labor involved in the production, implementation, maintenance and use of technology from the descriptions of their technologies. Moreover, she showed how the

[11] Breazeal et al. use *Joint Intention Theory* (Cohen & Levesque, 1990) to model the process of establishing mutual beliefs through 'communication acts' and 'conversational policies' (e.g. clarifications, elaborations, and confirmations). They enlist this theory to explicate how agents can have and maintain joint intentions, i.e. how they can share the same goal and execution plan. The conversational policies take the form of patterns of gestures, facial expressions and speech acts. Joint Intention Theory builds on various other theories, including Searle's *Speech Act Theory*.

[12] Harry Collins has explained that the automation or mechanization of tasks generally entails the creation of new responsibilities for humans, as they have to make good for the deficiencies of these automated technologies (Collins, 1990; Collins & Kusch, 1998). To make a machine work correctly humans have to perform a considerable amount of work in the form of 'repair'. Repair not only means modifying a machine to perform the appropriate action. Humans also interpret and adjust to the behavior of the machine to make it fit to practices, expectations and conceptual frameworks. I will return to Collins' work on automation in Chapter 4.

incompatibility between two modes of reasoning can lead to breakdown: the plan-based reasoning mode of the copying machine cannot accurately predict the outcome of the non-deterministic heuristic and situated action mode of the human. Various other critics have warned against the practical problems of trying to capture human behavior in the constraining frame of the computational and formal structures of systems designed to support humans (Collins, 1990; Friedman & Nissenbaum, 1997; Lanier, 1995). The unpredictability and complexity of human behavior makes it difficult to build systems that can accurately anticipate human preferences and actions. For this reason, 'interaction designer' Tom Erickson considered the use of the metaphor *context-aware* "deeply misleading" in descriptions of systems that detect and respond to features of their environments (Erickson, 1997, 2002). The phrase 'context-aware' as applied to computational systems, according to Erickson, can lead to misplaced expectations, breakdown and irritation when it is mistaken for the kind of context-awareness exhibited by humans.

Seemingly regardless of recurring objections, the idea that the interactive skills and the social knowledge of computers need to be leveled with those of humans remains a persistent theme in agent discourse. What makes it such a persuasive, but problematic, theme? The next sections will show that part of the answer lies within the decontextualized way in which visions of socially interactive computers frame the problem of human/computer interaction. A way to bring this framing into view is to consider the visions of future agents on a rhetorical level.

2.2 SEDUCTIVE VISIONS

Visions of proactive, intelligent and communicating artificial agents are part of a wider discourse concerning social change and technological development. The philosopher and STS scholar Harro van Lente points out that statements about future technological performance constitute language strategies to "mobilize attention, guide efforts and legitimate action" (2000, p. 43). The appeal to 'technological progress' as a necessary and inevitable force by optimistic visions of future technologies, he argues, is part of the rhetorical work of technologists, activists, politicians, firms, organizations, and other human actors to legitimize and mobilize support for their efforts, argument, project or cause. The abstract and ambiguous term 'technological progress' functions as, what van Lente calls in reference to the linguist Michael McGee, an *ideograph*.

McGee defines an *ideograph* as an abstract "ordinary language term" that represents a "collective commitment to a particular but equivocal and ill-defined normative goal" (quoted from van Lente, 2000, p. 45). An ideograph can be a forceful rhetorical instrument to mobilize support or excuse behavior without revealing the assumptions that underlie particular behaviors, beliefs, causes or projects. The term 'freedom', for instance, means so many different things that it almost has no real content. Nevertheless, it is a compelling notion as it represents a collectively constituted ideal that people should want, strive for and even die for.

As a result of their flexibility, ideographs, such as 'freedom', 'democracy' as well as 'technological progress' can be easily associated with and linked to a cluster of ideographs embedded in a shared cultural history. Through association, actors from different sides of a conflict can link an ideograph to historical cases with "accepted general lessons", and thus appeal to a shared sense of what is right and admirable or what should be denounced (van Lente, 2000, p. 46). Thus, advocates of the need for continuous innovation can point to the benefits of 'technological progress' in the past, such as the significance of Pasteur's inventions, or the gains in productivity as a result of computerization. In the same way, skeptics and Luddites can point to the evils that 'technological progress' has brought, such as global pollution and the atom bomb. Although the inevitability of technological progress is contested, as van Lente points out, the ideograph remains available and can be used as a rhetorical instrument "without any need for justification" (van Lente, 1993, p. 154).

The appeal to 'technological progress' lends visions of promising new technologies a persuasive force, van Lente notes. The ideograph 'technological progress' connotes ongoing evolution. Particularly in optimistic future visions, technological development is portrayed as an unstoppable incremental process of existing technologies being superseded by 'better' and 'enhanced' technologies. Against this background, technologies become 'obsolete', rather than fail, as new knowledge, technical expertise and technological resources expand the space of possibilities. Technological progress therefore necessitates the search for improvements to conventional technologies and demands action. It requires us to find new promising opportunities for innovation, because not "to have the next generation is to commit collective suicide" (p. 154). This conception of technological progress as an unstoppable evolutionary process supports an 'instrumental' way of thinking, where a

new technology "is necessary because the old is not good" (van Lente, 2000, p. 55). It is the answer to a problem that needs to be solved.

In agent discourse, an instrumental way of thinking is clearly visible. New developments make conventional computers obsolete or unsuited to deal with the current information jungle that humans are confronted with. These computers are the result of limited technological resources and incomplete knowledge. In the face of ongoing developments in computer science and in society, these limitations demand action to find new and 'improved' ways of building human-centered computational technology. Maes' vision of intelligent agents, for instance, is based on the idea that there is a growing problem with conventional computers. It is because computers become more and more integrated in daily life, that users, overloaded with information and work, need some form of (technological) assistance. She first points to the increasing complexity of computing technologies and the difficulties that novices or "untrained" users have with operating them. Additionally, she mentions her and her colleagues' dissatisfaction with the "time wasting" manner in which current tasks have to be dealt with by the user, such as "dealing with junk mail, scheduling and rescheduling meetings, searching for relevant information among heaps of irrelevant information, and browsing through lists of books and music and television programs in search of something interesting" (Maes, 1994b, p. 35). She then positions intelligent agents between the tasks and the user as a "radical" new style of human-computer interaction, where agents "hide the complexity of difficult tasks, they perform tasks on the users' behalf, they can train or teach the user, they help different users to collaborate, and they monitor events and procedures." (Maes, 1994a, p. 31). Maes thus frames the problem of current technologies in terms of their limited ability to appropriately support their human users. The task of agent researchers, then, is to find new ways that can relieve humans from the burdens imposed by ongoing technological development.

A narrative like the one expressed by Maes derives its strength both from the suggestion of necessary 'technological progress', as well as from the associations with a range of other ideographs, such as 'efficiency', 'user-friendliness', 'transparency', 'empowerment' and 'human-centeredness'. Ideographs that connect the development of socially interactive agents to ideals pertaining to social change and improvement add a seductive quality to visions of agents. The various accounts of agent advocates testify to an enthusiasm and ambition reminiscent of utopian visions of early eras. The philosopher Rein de Wilde points out

that current rhetoric about the promises of new technologies echoes themes of visions of ideal societies, which go back to the Renaissance (de Wilde, 2000). In analogy with earlier *political* and *religious* utopias, he argues, current techno-enthusiastic visions project images of a coming age that would see a return to peaceful, just, free and transparent societies.[13] This time around, however, it is not a political system or religion, but technology that offers the possibility of bringing us closer to a perfect society. Take, for instance, an excerpt of the website of the AmI project at the Frauenhofer Institute under the name CHIL:

> We aim to realize computer services that are delivered to people in an implicit, indirect and unobtrusive way. This will free people to interact with people and reposition machines to be in the background and - like electronic butlers - attempting to anticipate and serve people's needs. Computers in the Human Interaction Loop (CHIL) aims to introduce computers into a loop of humans interacting with humans, rather than condemning a human to operate in a loop of computers. This will give humans the most valuable gift: more time.
>
> *[emphasis in the original](CHIL project homepage, 2006)*

CHIL's website illustrates the more common image in agent discourse of a world in which electronic butlers will liberate humans from the arduous tasks of managing and navigating the increasingly computerized environment and the growing amounts of data and information that come with it. In this world agents offer accessibility to a happy stress-free life, where humans are free to act intuitively and do "the things of most interest to them" (Lieberman and Selker, 1999, p. 11). As these user-friendly technologies do not require expert knowledge or additional training to operate them, anyone can benefit from their advantages.

Narratives about the promises of technological advances with such a utopian flavor project seductive and persuasive visions of technologically enabled future worlds. They are seductive, de Wilde suggests, because they build on abstract ideographs that connect a number of previously conflicting ideals. He notes that various accounts of the promises of

[13] De Wilde refers to the ideas and images of future technological societies presented by "digital thinkers", such as Nicolas Negroponte, as *postmodern utopianism*. It is postmodern, de Wilde explains, because it differs from modernist utopianism in its emphasis on individualism and its embrace of capitalist traditions. Moreover, this form of utopianism conveys the expectation that new information and communication technologies will bring us beyond modernism. New technologies will dissolve the many dualistic distinctions that we inherited from modernism, including the distinctions between work and play, effort and relaxation, humans and technologies (de Wilde, 2000, p. 29).

increasingly intelligent technologies, including those of intelligent highways, intelligent cars as well as intelligent electronic assistants, regularly connect the ideograph intelligence to other ideographs such as safety, progress, transparency, efficiency, cost-savings, user-friendliness and even environmentally-friendliness. The techno-enthusiastic visions seem to suggest that the tensions that currently exist between these ideals can be overcome by the development of increasingly intelligent technologies.

Ideographs enable agent advocates to construct seductive and forceful visions of socially interactive agents. However, de Wilde reminds us that ideographic language is characterized by its abstract nature. When studied more closely in the context of use, the conflicts that the envisioned technologies generate - highlighted by the previously mentioned objections to socially interactive agents - present themselves. These conflicts indicate the presence of normative choices that underlie the use of ideographs. The ambition to develop technologies that anticipate and attend to user preferences, de Wilde notes, imposes certain constraints. He draws attention to the opposite side of the coin of 'user-friendly' human/technology interaction. The emphasis on the need for 'effortless' interaction with information and communication technologies (ICT) also entails that there is less demand for humans that have the critical and intellectual skills to understand how these technologies function (2000, p. 132). As computers become increasingly smart and complex in order to support the human user, only those with expert knowledge will have privileged access to the complex networks that agents hide.[14] The normative choice that this trade-off poses disappears from view as a result of the decontexualizated descriptions of the envisioned technologies.

Considering agent visions in terms of language strategies encourages us to further explore the assumptions underlying the promises of the development of agents in order to expose the normative choices that they imply. What do agent researchers mean when they speak of improving 'user-friendliness'? How do they frame the problem of human/technology interaction and in what sense are socially interactive

[14] Sherry Turkle has expressed concern about the tendency to hide the complexity of computers behind interfaces, as it limits the need for understanding how computers work (Turkle, 2005). Whereas in the 1980s computer education for children involved teaching them about programming and how algorithms work, today it is more focused on using appliances rather than understanding the structure behind them. Although knowing how to operate a computer can empower people in certain areas, not knowing how the computer operates makes them dependent on other people's creations.

agents the solution to this problem? What does this framing hide? In the next section, I will address these questions by taking a closer look at how agent researchers go from identifying the 'flawed', 'archaic' or 'outdated' nature of conventional computer technology, to advocating the need for the development of autonomous, socially capable and intelligent electronic entities.

2.3 A NARROW VIEW

In agent discourse, the interpretations of elusive terms, such as 'user-friendliness' or 'human-centeredness', is based on the key assumption that 'effort', in particular in the form of cognitive work performed by humans, equals 'obtrusive', 'rigid', 'time-consuming', 'unnatural', and 'burdensome'. The way we experience conventional computational technologies, such as desktop computers or mobile phones, is portrayed as unsatisfactory as a result of the (mostly cognitive) demands placed on the user by the interface of the device or system.[15]

The ambition to reduce the effort involved in operating technology underlies Lieberman and Selker's call for the development of context-sensitive interface agents. They consider the inability of computers to take account of the circumstances under which actions are performed as a serious flaw of contemporary computing devices. Acting "exactly the same regardless of when and where and who you are, whether you are new to it or have used it in the past, whether you are a beginner or an expert, whether you are using it alone or with friends" makes systems "brittle" (Lieberman & Selker, 2000, p. 618). They criticize what they call the "black box" view in software design approaches. This view, they

[15] Most computers these days come with some kind of Graphical User Interface (GUI) often designed based upon some form of *Direct Manipulation* (DM), an interface design style introduced by Ben Schneiderman. DM provides a set of design principles, which describe an interface as consisting of continuous representations of objects that can be manipulated through actions loosely based on real-world physical metaphors (Schneiderman, 1983). Through actions on the object representations the user can manipulate the formal computers structures that are connected to the interface. Examples of DM interface elements include windows, mouse pointers and icons, but they can also refer to sonic or tactile control feedback loops. To operate these computer systems, humans have to conform to the constraints of the interface, which can entail that technology demands expert knowledge from the human that has to be acquired through a learning process. Although Schneiderman introduced his design principle as yielding 'natural' and 'intuitive' interfaces, several agent advocates present a DM style of operating a computer as 'obtrusive', 'rigid', 'time-consuming', 'unnatural', and 'burdensome'.

argue, considers a computer program to be a *context-free* abstraction. Like with mathematical functions, in this software design approach consistency is a key requirement. Given a particular input, software should produce the correct output irrespective of the context. Lieberman and Selker argue that this approach leaves it up to the user to translate the context to the computer by providing explicit input and interpreting the subsequent output in light of the context. Moreover, the possibilities for interaction are constrained by what the formal model of the computer and its context-free representations allow. Thus, when I use software for the first time, I could read the manual or experiment in order to figure out how I can get the computer to do what I want it to do. The computer obediently waits for my instructions to execute them, while I have to conform to abstract representations of computer processes in order to convey what I need the computer to do. Furthermore, my laptop is not aware of the circumstances under which I am using it. It will act in the same way regardless of whether I am in the office or giving a PowerPoint presentation in front of a room full of people. These features of most soft- and hardware today, Lieberman and Selker claim, are burdensome and occupy the human user unnecessarily.

From a perspective that considers cognitive effort to be problematic, the 'unnatural' actions required of the user to operate conventional computer devices indicate that the technology does not 'fit' well with human activity. The amount of effort and the time it takes to master operating a computer system as well as the time spent on instructing it serve as a measure for the level of fit. This understanding of what is wrong with current day computers reflects a normative assumption: development should go from complex technology-driven opaque technologies, only to be used by those initiated in the mystical world of computers, to technology for the "rest of us" (Breazeal, 2002). One conclusion that, for instance, advocates of both UC and AmI draw is that technologies should be 'self-explanatory' and eliminate the steep learning curve associated with current computing technologies. According to Mark Weiser, the specialized skills required to operate a computer are in no direct relation to the task that is being performed, and make it the undeserved center of attention (Weiser, 1991). From this perspective, humans and technologies are two separate mismatched entities. A gap exists between humans and technologies (as well as between groups of humans) that needs to be bridged.

Against the background of a perceived flawed interaction with obtrusive, irresponsive and complex technology, the metaphors of 'social

interaction' and 'communication' support a conceptual framework with which to formulate an alternative, 'better' kind of interaction. In this framework, the gap between humans and technologies can be explained in terms of the limited means of interaction. A re-conceptualization of technology as 'partner' to communicate with - rather than 'tool' - is then easily imagined. Reinforced by a connected set of ideographs, such as 'user-friendliness', 'intuitiveness' and 'progress', agent advocates structure their visions around the idea that human/technology interaction needs to be improved by leveling the competences of artificial agents with those of humans. 'Intuitive' and 'natural' human communication and social interaction serve as a reference point; as something to aspire to when developing new technologies in contrast to tools. An intelligent agent endowed with social knowledge and communicative skills should be able to take over a larger part of the burden of translating and interpreting context-dependent information and adapt to individual users.

Breazeal justifies her work on the Leonardo Robot, based on the argument that human-like communication enables a more 'natural' interaction between humans and technologies. Although, Leonardo has a high cuteness-factor and appeals strongly to one's imagination, the ultimate aim of Breazeal's group extends further than building adorable animate toys. Besides providing an empirical test bed to experiment with theories about human cognition and social learning, she claims that humanoid robots that are able to elicit social behavior from humans towards them offer a number of advantages for future applications:

> First, people would find them more enjoyable, and would thus feel more competent. Second, communicating with them would not require any additional training since humans are already experts in social interaction. Third, if the robot could engage in various forms of social learning (imitation, emulation, tutelage, etc), it would be easier for the user to teach new tasks. Ideally, the user could teach the robot just as one would teach another person.
>
> *(Breazeal, 2002, p. 16)*

Breazeal assumes that social interaction is the primary and most 'natural' mode of action and communication for humans in confrontation with increasingly complex technology. This, she holds is the result of evolution "hardwiring" "innate mechanisms" in the human brain that enable them to act in a social manner (p. 15). Given these inherent skills, the simulated 'natural language' and social cues performed by an artificial agent should guide human users in their dealings with it.

Building on an evolutionary narrative of technological progress, agent researchers, such as Breazeal, characterize the current gap between humans and technology as representing an early stage in technological progress. Current technologies have not evolved far enough to exploit the full potential of communication. Improving this interaction is then merely a matter of complexity and ongoing development. This framing of the problem and its solution conveys particular reductionist conceptions of humans and human agency, which are reminiscent of the representations of humans and human action in AI rhetoric criticized by social scientists and feminist scholars (Suchman, 2008).

Despite the centrality in agent discourse of the human as goal to aspire to, the idea of interfaces as agents leaves a remarkably small role for human agency. Take, for example, the reasoning behind the development of the intelligent tutoring system called AutoTutor (Graesser, Person et al., 2001; Graesser, VanLehn et al., 2001). This system, developed by the Institute for Intelligent Systems led by Art Graesser at the University of Memphis, is equipped with an animated pedagogical agent to give an output that is more 'intuitive' and 'natural' for the user and makes the interaction more attractive. This moving and talking graphical face acts like a "conversational partner" and delivers dialogue moves with synthesized speech, intonation, facial expressions, and gestures (Graesser, VanLehn et al., 2001, p. 41).[16] Graesser's group starts from the premise that in order to stimulate a student to acquire a "deep understanding" of the material she should learn through a natural language dialogue (p. 45). The 'tutor' will ask questions, encourage the student to elaborate her answer through positive and negative feedback, prompting or hinting. Greasser and his colleagues claim that in this way the system encourages the student to articulate long answers that show a "deep reasoning", instead of reciting "shallow knowledge" (p. 41). They assume that the computer manipulates the student's behavior to affect the efficacy of the learning process and that this process can be formalized. However, in doing so, they 'design' a particularly limited conception of student users. They portray students as rather passive sponges that learn effectively only when directed towards the 'right' way of absorbing information.

In the context of use, the complexities of the role of human agency in shaping the interactions between humans and technologies come back into view. Du Boulay et al., for instance, discuss the *plausibility problem*

[16] The current version of AutoTutor is configured to teach in Newtonian physics and Computer literacy.

that became apparent in their empirical studies of two types of intelligent tutoring system (du Boulay et al. 1999). This problem emerges when expectations of a system's abilities are not in agreement with its actual capabilities. Du Boulay et al. mention problems such as students refusing to accept the pedagogical decisions of the system, and students underestimating the capabilities of the system. The plausibility problem underlines the role of human agency in shaping a successful interaction analysis, which literature in STS has also drawn attention to (Akrich, 1992; Oudshoorn & Pinch, 2003). The one sided-focus on the promises of developments in agent technology of Greasser and his team pushes the role of human agency to the background, as it isolates the technology from the contexts in which it becomes connected to humans. As result, they seem to transfer the agency of humans to the agents (Suchman, 1998).

Agent advocates often reinforce the portrayal of the human user losing out in terms of agency as compared to smart intelligent electronic assistants, by isolating the concept of the human from a historical and cultural context. They supplement the evolutionary theme that characterizes 'technological progress' rhetoric with a Kurzweilian evolutionary theme, in which human evolution has virtually come to a halt and where technological evolution is exponentially progressing. The increasingly complex digital environment is no longer comprehensible for humans, because human evolution cannot keep up with the fast pace of technological evolution.

The wide adoption of ideas expressed in the work of Byron Reeves and Clifford Nass, in much of the above discussed research, exemplifies the pervasiveness of the idea that human evolution is trailing behind (Bartneck, 2006; Breazeal, 2002; Fong et al., 2003; Markopoulos et al., 2005). Reeves and Nass are responsible for the theory of Computer As Social Actors (CASA theory), which has inspired a range of applications that are built to be 'social actors', such as Microsoft's Clippy, as well as Breazeal's humanoid robots. In their seminal book *The Media Equation* (1996), Reeves and Nass discuss a number of experiments, which suggest that people respond to social cues expressed by computers in a similar way as they would when humans exhibit these cues.[17] They report, for

[17] Reeves' and Nass' idea of social responses is typified by 'social or natural rules' that summarize findings taken from social science about how people respond to each other. For example one rule they examined in their experiment is: 'People like to be praised by other people even if this praise is undeserved'. Their methodological approach consist of series of steps to draw such 'rules' from social science research (mostly social psychological research) and test these rules on human/technology configurations.

instance, that participants in one of their experiments gave significantly more positive responses when queried by a computer about its own performance, whereas they were inclined to give more honest evaluations when queried through independent media. Reeves and Nass conclude from these observations that people are subconsciously polite to computers.

Based on their experiments, Reeves and Nass claim that "individuals' interactions with computers, television, and new media are *fundamentally social and natural,* just like interactions in real life" (1996, p. 5). Humans can be well aware of the inanimate nature of the application and still respond socially.[18] However, humans simply cannot help responding in a social manner to certain artifacts, even if they are aware upon reflection that the social behavior of the computer or media is not 'real'. This behavior, they argue, is the result of our slow evolution with regards to the growing presence of technology in our daily lives. We are simply not calibrated to automatically treat inanimate objects according to a different protocol. It is only after rational (delayed) reflection that humans are capable of making the distinction between animate and inanimate in terms of their response to media.

This conception of the developing relationship between humans and technologies too easily ignores the flexibility and adaptivity of the human user. Ironically, it is Mark Weiser, the father of Ubiquitous Computing, who provides us with an elucidating example of the feats of human adaptivity. To argue for the idea of unobtrusive technology, Weiser gives 'writing' as an example of a profound technology that is ubiquitous and that does not require active attention. "Today this technology is ubiquitous in industrialized countries. [. . .] The constant background presence of these products of 'literacy technology' does not require active attention, but the information to be conveyed is ready for use at a glance" (1991, p. 94). This is in contrast to silicon-based information technology which, according to Weiser, "remains largely in a world of its own" (ibid). However, reading and writing are not such natural intuitive skills as Weiser makes them out to be. Proficiency in reading to the extent that it no longer requires 'active attention' is a skill that can take many years to master. People suffering from dyslexia are very much aware of the effort required to operate 'literacy technology' effectively. Generally most profound technologies become profound after a process

[18] Reeves and Nass note that they do not claim that humans have a tendency to *anthropomorphize* behavior of media, which they define as having "the mistaken belief that inanimate objects are human" (1996, p. 10).

of familiarization. Extended use and appropriation by human users can let the object disappear from their conscious minds. Even the personal computer is not experienced as an obtrusive device by everyone. For some, using a Word processor feels more natural than writing with pen and paper.[19] Whether this is a positive development is subject to debate.

The technology-biased evolutionary conception of future human/agent relationships overlooks what can best be described as the *co-evolutionary* processes that produced many of the human/technology relationships that exist today.[20] Not only do humans learn to use technologies throughout their individual lives, the connections between humans and technologies constitute relationships that evolve over generations, where both human beings and technologies become more finely attuned to each other.[21] This co-evolutionary process ensures that these relationships become strongly embedded and solidified in social practices. The ecological psychologist and sympathetic critic of artificial agents, John Pickering, remarks in this respect: "Human beings develop within an envelope of skilled practices and the material artefacts [*sic*] associated with them. Together these constitute a self-replicating system that leave a permanent trace, both within the environment and with the body" (Pickering, 1997).

Thus when AmI and UC proponents speak of replacing the keyboard, windows and pointer metaphor with more 'natural', 'intuitive' interaction strategies, they are assuming that the operating users are hindered by and uncomfortable with these conventional strategies. Yet this assumption is not self-evident, as children now growing up with a computer as a pervasive element of their environment experience interacting with technology differently (Docampo Rama, 2001; Lauwaert, 2007). Typing on a keyboard and using windows, icons and pointers might seem complicated to people used to writing on typewriters, but children growing up with these tools might find them a more natural way of

[19] The example of writing also illustrates that social interaction is only one mode of humans interaction with the environment (Hutchins, 1995). Shifting the focus can provide alternatives to the metaphor of social interaction (Dourish, 2001). The project of Tangible Computing, for instance, places the emphasis on embodied physical interactions with the environment as a 'natural' form of interaction(Ishii & Ullmer, 1997). This line of research aims to exploit human physical and tactile skills.

[20] Co-evolution (also referred to co-production or co-constitution) is a central concept in technology studies, where it is used to capture the key assumption that technology and society mutually shape each other over time (Bijker & Law, 1992).

[21] Henry Petroski provides an enlightening and detailed account of the continuous development through history of various mundane and 'useful' artifacts, such as forks, knives and paperclips, and their associated human skills (Petrosky, 1992).

dealing with electronic tools. Metaphors or models that seem natural for one generation are determined in part by culture and experience (Turkle, 1984). This culture, in turn, is dynamically co-evolving with the technology that is integrated and adopted by it. The longing for an augmented environment that resembles that of the home before the introduction of the personal computer seems to be motivated by the nostalgia of a generation that grew up without it.

As mentioned, a recurrent objection to the application of the metaphor of social interaction to human/technology relationships is that it relies on a reductionist conception of humans and human agency. However, to understand why the problematic aspects of these conceptions do not deter agent researchers from advocating their visions, we have to look towards the multiple conflicting concepts of humans and agents that these visions build on. In leveling agent and human competencies with regard to their interaction or communication, advocates of socially interactive agents not only put forward a new role for these agents to play; they recast the human role as well. They construct two separate abstract notions of the human as ideal and the human as 'user'. On the one hand, the idea of being human serves as a point of reference for the formulation of a higher goal to be achieved. This idea attributes properties to humans, such as intelligence, adaptive and flexible behavior, learning and social communication, as opposed to the deterministic reasoning capacities and rigidity of conventional computers. In order to be the effective assistants their advocates desire them to be, agents should be equipped with similar human faculties. Yet on the other hand, the human as 'user' in the visions discussed is not a dynamic, continuously configured entity, but a stabilized, isolated notion stripped of autonomy and defined by formalizable patterns of behavior.[22] The human as user is a fixed entity that needs to be accommodated and is no longer capable of readjusting to and operating within quickly expanding digital networks without the assistance of computational partners. It is because the discourse disconnects these two notions, by abstracting the idea of the human from its context, that the inherent conflicts between them do not present themselves as problematic.

[22] Critics have highlighted other problematic abstractions in the representations of the user in the design of computer technology. Feminist scholars of technology, for instance, have argued that designers tend to construct a narrow conception of the future user: very often this is a young, highly educated white male. See, for instance, Oudshoorn et al. (2004).

The conceptions of socially interactive agents show an ambiguity that is directly related to the conflicting notions of humans. Agent advocates move the conception of computer programs and robots away from instruments to extend and amplify human cognitive and physical capabilities, towards computational assistants or partners that operate jointly, in negotiation with, or on behalf of, humans. In constructing their visions, they represent artificial agents as independent entities modeled after an idealized notion of humans. Agents learn, communicate and pursue their own goals, without suffering from the problematic aspects of human social interaction, such as misunderstanding and conflicting interests. At the same time, agent advocates never really leave the discourse of tools, as they continue to focus on what technologies should do for humans. When the envisioned agents are discussed in terms of why they are being developed, they are represented as instruments that serve as extensions of the abilities of humans. The agents operate within the narrow constraints of the specification of their envisioned function. In her reflection on intelligent agents Suchman identifies a similar tension (2003). She notes that: "there is a deep and enduring ambivalence [. . .] inherent in the image of the agent: on the one hand, the agent as faithful representative, on the other hand, the agent as autonomous, self-directed, and therefore able to pursue its own agenda" (p. 41).

As rhetorical instruments, the ideographs and metaphorical concepts in visions centered on the idea of moving technologies closer to humans remain abstract ideals isolated from the contexts of practice in which they gain various meanings. As such, they help build an image of a human/agent relationship in which competencies and skills can be transferred and delegated from one entity to the next. Veiled by evolutionary and utopian themes, the abstract notions of humans and agents make it possible to construct an enticing narrative in which technologies develop from instruments meant to extend and complement human cognitive and physical capabilities, towards separate entities that will operate jointly and in negotiation with humans. Nevertheless, the decontextualized concepts of humans and agents generate a contentious and problematic vision of future human/technology relationships. They frame the discourse, such that it preferences a single vision of a flattened two-dimensional relationship between two equal entities. This framing obscures the view from the normative choices in configuring the connections between humans and technologies. Visions of intelligent computational assistants present a trade-off between on the one hand

minimizing the skills and cognitive effort required of the human user to perform particular tasks, and on the other bolstering her control and responsibility over these tasks. By delegating a larger part of decision making to opaque complex computer technologies the user is limited in her abilities to question the way an action is performed (van den Hoven, 2002). Although I find Amazon.com's book recommender a tremendous help, I have limited knowledge and little control over the way it constructs a profile of my perceived preferences (or what it does with this profile). I therefore have limited means to question its recommendations. Agents that would, on the basis of their perceptions of me, take over a larger part of the responsibility of interacting with and deciding for me increase this dependency even further. In addition, agent advocates generally fail to mention the possible consequences of their envisioned systems with regard to such issues as privacy and security. Technologies 'for the rest of us' in the form of socially interactive agents that act on our behalf, come at the price of introducing new problematic dependencies.

A pre-occupation with the leveling of humans and technologies leaves little room to explore the potential of other processes that produce human/technology activity patterns in existing relationships. In fact, the discussed visions dismiss these relationships, by characterizing them as "unnatural" or "burdensome". Suggesting a shift away from the tool metaphor, as a result, draws the attention away from the complementary roles of technological artifacts (as distinct objects) in enabling particular kinds of actions. The next section shows how a more contextualized perspective on multiple dimensions of the connections between humans and technologies can broaden the view of the landscape of possible human/technology configurations.

2.4 SPACE OF POSSIBILITIES

In 1987 Terry Winograd and Fernando Flores wrote *Understanding Computer and Cognition*, in which they criticized the basic premise of AI that computers could and should be made to be like humans. Their critique of classic AI now seems somewhat outdated, as it attacks the notion of building isolated rational symbol processors, something that has become far less prominent. Nevertheless, they make a subtle point about the position of computers in relation to humans that is still valid: computers are differently situated in activities as compared to humans when it comes to things like decision making and communication.

Winograd and Flores claimed that the role of the computer is not to mimic humans, but to act as a 'linguistic tool' (1987). They conceive of computers as a means to facilitate and support communicative actions performed by humans, instead of as a participant.

Language, for Winograd and Flores, is a decidedly social activity. Giving meaning to language "is rooted in our participation in a society and a tradition" (p. 61). In their view, human communication is an essential mechanism for humans to perform and coordinate actions with other humans. It serves a variety of functions besides conveying information, such as aligning interests and goals, sharing beliefs and ideas, maintaining social relations, but also engaging in or perhaps resolving conflicts. Winograd and Flores claim that technology, and in particular computers, are part of communication mechanisms, but differently so. The design of computers, they hold, is rooted in a modern rationalistic tradition that thinks in terms of information and representation being transported from one entity to the next. This tradition views technology as isolated and ignores the overall system in which technologies are embedded. It takes language as a "carrier of information", rather than a mechanism for negotiating meaning that continuously and dynamically develops in human social practices. Winograd and Flores maintain, therefore, that computers cannot engage in communication in the same way as humans do. Nevertheless, they can facilitate and direct communication. "Their power as tools for linguistic action derives from their ability to manipulate formal tokens of the kind that constitute the structural element of languages" (p. 76). Computers can support human collaboration, by enabling humans to coordinate actions. For this reason, they say that the idea of making machines like humans is misguided, as it is in other domains that the computer can make better and more contributions.

Winograd and Flores characterize communication between humans in terms of the obligations, commitments, rights and expectations that humans use to coordinate their actions. From their perspective, as long as computers cannot be part of this, they cannot take the position of humans. However, this is exactly what agent enthusiasts, like Breazeal, aim to achieve. Breazeal's ultimate objective is to make agents take part in human culture, as equal partners in communication. The crucial point of Winograd and Flores' analysis is not so much to question the viability of this ambition. Rather, they show that because technology plays a different role it makes particular kinds of human action possible. In their role as linguistic tools, these technologies *mediate* the relationship

between humans and the world.[23] They affect and transform human communication, as the characteristics of this tool shape and form human actions and perceptions.

The analysis of Winograd and Flores demonstrates that the differences between humans and technology can complement each other, rather than frustrate their interaction. The emphasis on the leveling of humans and agents in the rhetoric on artificial agents creates a blind spot for these complementary roles. The abstract notions of both humans and technologies obscure the view of the characteristics of the particular interactions between humans and technologies within specific contexts. Winograd and Flores adopt a relational perspective in their analysis of interactions between humans and technologies that shifts the focus to these characteristics. It highlights that the asymmetries between humans and technologies make particular activity patterns possible. We should, however, be careful not to get caught in the same trap by narrowly focusing on the instrumentality of computers as linguistic tools. Winograd's and Flores' approach to analyzing the role of both humans and technologies in their interactions offers the means to consider a wider range of possible human/technology configurations.

A re-evaluation of the asymmetries between humans and technologies brings into view how the properties of particular connections between humans and technologies can affect actions. A conversation between friends over the phone takes a different form than, say a conversation on the MSN messenger. A telephone conversation allows for a more subtle expression of emotion or affection, than a typed response on MSN. An advantage of talking to a friend over the MSN messenger is that it does not require continuous attention. Donald Norman appropriated the term *affordances,* first introduced by the psychologist James Gibson, to capture the different properties of the relationships between particular humans and particular technological devices (Gibson, 1986; Norman, 1993). An affordance is not a property of a device; rather it is a property of the

[23] The way I use the term *mediation* here refers to notion of *technological mediation*, as described by the philosopher Peter Paul Verbeek (Verbeek, 2000). Verbeek uses the concept of technological mediation to capture the role of technology in shaping the relationship between humans and the world. According to Verbeek, technology transforms our actions and shapes our experiences and perceptions of the world. He is particularly concerned with how technology mediates our being in the world. Verbeek's application of the terms *mediator* or *medium* to technological artifacts differs from Winograd's and Flores' use of the term *medium.* They follow Maturana and Varela, who used the term to refer to the space in which an entity exists, which includes the environment as well as the entity.

relationship that holds between a device and an individual operating or acting on the device. Affordances are both enabling and constraining. A particular technological device provides the possibility to perform a certain range of actions. For example, letters as means of communication afford reflective contemplation, because, unlike an MSN conversation, a conversation via letters does not take-place in real-time. Television and radio "afford one-way communication from performer to audience, but they do not afford communication in the reverse direction" (Norman, 1999, p. 125). Reversing the one-way communication of the radio would afford a more interactive form of communicating, but it will also restrict the range of actions a human can perform: operating interactive devices like a telephone limits the possibility to do two things at once, as it demands a certain level of attention.

The concept of affordances highlights the 'active' role of technological devices in shaping the human/technology relationship. This active role comes about in the interplay with the cognitive and physical competencies of the human user under particular circumstances. The actions that a device affords, as Norman notes, are particular to the individual acting on the device. For instance, a heavy stone does not afford a little child to throw it, but it might afford me to do so. To describe the affordances of a particular technological device, we therefore have to look beyond the structural properties or functionalities of the device and incorporate the expectations, knowledge, and experience of the human using the device as well as its context of use.[24] Norman's concept of affordance highlights the context-dependent, inextricable connections between humans and technologies that are hidden from view in abstract descriptions of isolated humans or technologies. Although the design of the technology provides a set of conditions for action, the form and meaning of these actions are the result of the configuration of particular humans and technologies in particular contexts.[25]

---

[24] Norman points out that 'real' affordances (i.e. does the stone actually afford me to lift it) are not nearly as important as *perceived affordances* (Norman, 1999, p. 123). The human has to be able to recognize the affordances. According to Norman, *perceived affordances* convey to the user what actions can be taken and how the object should be used. They are mostly about conventions and what the user knows or can perceive. Thus, a child-proof screw top on a bottle of cleaning chemicals only affords those that know the trick to open it. Perceived affordances explain why it is that some technologies are 'easier' or 'more attractive' to use than others.

[25] Bruno Latour and Madeline Akrich make a similar point in their analyses of technological scripts (Akrich, 1992; Latour, 1992). In contrast to Latour and Akrich,

From the point of view provided by the notion of affordances, the combination of distinct qualities of both humans and technologies generates particular activity patterns. This observation casts a different light on the assumptions that underlie visions of socially interactive agents. It allows for a re-evaluation and reconsideration of concepts such as 'seamlessness' and 'user-friendliness'. For instance, the frictions that the use of technology might yield can work to enable a particular kind of action that would not be possible if it did not exist. It is because we have to browse through libraries that we can find new sources of information. In the protracted activity of drawing a sketch of a landscape, the viewer can gain a deeper appreciation of the panorama, than capturing the moment with the quickness of a camera (Norman, 1999). Effortless is not always equal to human-friendliness. A model of human/technology interaction centered on the idea of minimizing cognitive efforts thus entails a normative choice concerning the nature of an activity. In different contexts this choice might be taken differently.

A focus on the context-dependent interdependencies between human and technologies reveals a broader spectrum of possible configurations, in comparison to the narrow view of agents as 'partners' or 'invisible interfaces'. Moreover, approaching these interdependencies from different perspectives encourages a further exploration of the various factors that affect them, such as human experience.[26] The philosopher Don Ihde talks of an *alterity* relationship to describe how humans and machines can relate in a positive or *presentential* sense *to* or *with* "technologies-as-other" (2003).[27] Like my own experience with the Spatial Sounds installation, the technology's 'objectness' becomes an *otherness* when it seems to have a life of its own. Using the technology becomes *interacting with* the technology, where the interaction becomes a dialogue or exchange.[28] Experiencing a 'spinning top' (Ihde's example) or a

---

Norman adopts a perspective that emphasizes the cognitive and phenomenological aspects of individual humans relating to technological artifacts. Latour and Akrich instead put the focus of analysis on the larger heterogeneous networks that are shaped by and shape technological artifacts. For them, conventions and background knowledge cannot be taken as a given, as they are elements in these heterogeneous networks.

[26] Verbeek distinguishes between the praxis and perception perspective on the relationship between humans and the world mediated by technology (2006). The praxis perspective highlights the role of technology in mediating human action and the perception perspective focuses on the mediation of human experience.

[27] In contrast to a negative sense where the technology derives its objectness from break-down.

[28] Ihde points out that otherness in the case of technologies can only be *quasi-otherness*. To illustrate what makes technological *otherness* a *quasi-otherness*, Ihde compares a car to a

spinning mechanical arm as *other* allows me to attribute some form of intentional agency to it. In a similar way, it makes sense to engage in communication with automata on our early encounters with them and experience them as *other*. This experience does not present itself arbitrarily; it is not unconditional. My encounter with the Spatial Sound installation took place within an entertainment setting, in which certain constraints, such as efficiency or professional performance, are absent. This context allowed me to suspend my disbelief and accept the technology as other. A laptop that would present itself as other, would easily lead to breakdown in the interaction. Leveling humans and agents in terms of their communication skills, in this case, changes the activity and places different demands on the human as well as on the context.

The experience of encountering an object or an artifact that invokes the feeling that it has a life of its own provides an enticing metaphor. As such, it appears throughout the history of humans and machine (Franchi & Güzeldere, 2005). However, models of human/technology interaction based on this metaphor, as well as models based on the idea of invisible agents, present a limited set of possible forms of interaction. As part of his phenomenology of technology, Ihde identifies a spectrum of types of "existential relationships" between humans, technology and the world. Drawing on Heidegger's idea of "ready-to-hand", he describes how when we use certain technologies we experience the external world *through* technology. The technology becomes an 'invisible' medium through which the external world reaches our perception. We can perceive the world through a pair of spectacles without being consciously aware of the presence of this information transforming device. In contrast to this embodied type of relationship, when I read the temperature on the display of a thermometer, I do not have direct access to the world. The thermometer is the object of my perceptual focus. However, it is not the object of my conscious consideration; I am interested in the temperature outside. In such a *hermeneutic* relationship, as Ihde calls it, the thermometer is the object of my perceptual focus, because it represents something about the external world. A central heating system retracts even further from our conscious experience of it.

---

horse. He identifies the horse's ability to exist without human intervention and its potential for disobeying as elements that allow for it to be experienced as other rather than mere object. According to Ihde we can relate to a car and interact with it, but it can only partially be attributed otherness. It lacks the flexibility and independence of the horse and remains under human control. The resistance of a horse is more than a mechanical lack of response – the response is more than malfunction, it is *dis*obedience (Ihde, 2003).

Our relationship to this device, Ihde calls a *background* relationship. Ihde's different "existential relationships" illustrate how we can experience technologies in different ways and consequently conceive of them as positioned differently in relation to ourselves. His categories not only highlight the multiplicity of human/technology configurations, they also reintroduce human experience as a variable that affects and is affected by the nature of the configuration and the context in which it takes shape.

The contextualized and relational perspectives discussed in this section highlight multiple dimensions of the human/technology relationship. They reveal a broad spectrum of possible configurations - and thus also a wider range of design choices - in comparison to the visions of agents as partners or invisible interfaces. The envisioned leveling of humans and technology, exemplified in visions based on the metaphor of socially interactive agents, represents an abstract conceptualization of only a small sub-set of these configurations. It highlights particular idealized features of human behavior, while it hides context-specific characteristics of interactions between humans and technologies. The main problem of building agents to be more like humans then is not so much the viability of this ambition; rather it is the narrow view that visions on artificial agents present. What we do with technologies and what technologies do with us, cannot be captured by a single metaphor for human/technology relationship. The danger of accepting a single abstract model of human/technology relationships as the best possible overall solution to interface problems is that it inhibits discussions on what technologies can and should do in relation to humans.

2.5 CONCLUSION

Social interaction and human-like communication provide attractive metaphors to explore new ways of connecting humans and technologies through the interface. They offer a framework in which to consider computers as 'more than tools', as entities that interact in a 'natural' and 'intuitive' way with humans. In this chapter I have considered agent visions that cultivate this metaphor to examine why the idea of bridging the gap between humans and technologies continues to be a persuasive, but problematic feature of discourse on future technologies. A recurrent theme in the computer science discipline is the presupposition that in order to improve human/technology interaction, the (social) competences of computers have to be leveled with human abilities. Critics have

repeatedly highlighted the limitations and risks of modeling human/technology interaction after human-to-human interaction. They have challenged the reductionist conceptions of humans and technologies, which constitute models and theories of socially interactive technologies, and have pointed to consequences of these conceptions. Despite these recurrent objections, agent researchers continue to pursue and argue for the development of computer technologies that move increasingly closer to humans. As the analysis in this chapter shows, a reason for the persistence of this ambition can be found in the way that metaphorical concepts are used to frame perceived problems of human/technology interaction.

Visions of socially interactive agents demonstrate a seductive and forceful rhetoric that presents the development of these types of agents, as the necessary and natural next step towards an optimal interaction between humans and technologies. Using metaphors like social interaction and human communication, agent advocates frame the problem of human/technology interaction in terms of flawed communication. The solution of leveling the competences of humans and technologies then seems like a promising and necessary solution, when couched in narratives of technological progress and supported by ideographs such as user-friendliness and effortlessness. The rhetorical strategies mask ambiguous conceptions of humans, technologies and of the connections between them, and distract the attention from the unfavorable implications that the envisioned agents might have. As a result of the decontextualized nature of the discussed visions, the inconsistencies within these visions do not immediately present themselves as problematic. In fact, the hidden ambiguities are part of the appeal of these visions. They make it possible to sketch an image of autonomous entities that work on our behalf and seamlessly adjust to our needs and preferences.

To engage in informed debates about the promises of new technologies and the choices that these technologies present we need to look beyond ideographic language. The visions discussed in this chapter on their own provide an inadequate basis for broader debates about future technologies as well as for the development of agent technologies. They hide underlying normative assumptions and lead to a preoccupation with bridging the gap between humans and technologies. Broader contextualized and relational perspectives on the connections between humans and technologies showed a variety of contingent factors that affect how these connections take shape. Expectations, human skills, knowledge and experience as well as the physical and social environment affect how

technologies relate to humans. From this point of view, agent visions provide a narrow and contentious view of possible human/technology configurations. 'User-friendly', 'intuitive', and 'effort' are equivocal and contestable concepts. Their use reflects normative assumptions about how technologies should relate to humans, which become apparent when considered within use practices. We can for instance ask whether we want interfaces to hide the complexity of computer systems. Investing more in teaching users how a technology works can be a preferable strategy over striving to minimize the time it takes for users to learn how to work with the technology.

Empirical studies in fields like STS and HCI have demonstrated that optimizing the interaction between humans and computers involves a number of questions that can only be answered through contextualized analyses of human/technology relationships (Bijker 1995, Winograd, 2006). What constitutes 'better interaction'? What kind of knowledge and skills do humans have to have to work with the device? To what extent and which human actors should be in control of the technology? What tasks should be delegated to computer systems? What norms and values should be reflected in the design and how does this design affect human action? Insights from empirical studies of human/technology relationships are therefore a valuable contribution to the exploration and evaluation of new models for human/technology interaction. Such studies allow us to challenge and examine the assumptions underlying agent visions and explore alternative routes to connecting humans and technologies.

Considering the possibilities and limitations of innovative approaches to modeling human/technology interaction requires a broader focus than the narrow view that the enticing, but abstract visions of socially interactive agents present. A first step towards this end is to acknowledge that agent visions are constituted by metaphorical concepts. An awareness of the metaphorical nature of these visions invites an analysis of the way in which metaphors help to frame problems and how this framing supports particular conceptions of humans and technologies. A critical interrogation of this framing can help AI researchers explore and experiment with different ways of conceptualizing human/technology relationships. Social interaction and communication are only two of the available metaphors to model these relationships. Choosing another metaphor or interpreting metaphors differently sheds a different light on the problem of interaction, offering alternative design possibilities and, thus, different choices. The various metaphors for human/technology

interaction are, therefore, best conceived of as possible options within a collection of conceptual tools, including 'direct manipulation' and 'linguistic tools'.

A contextualized analysis also means that we have to consider the various discourses in which agent visions appear. The focus of the discussion in this chapter has been on the rhetoric about the promises of bridging the gap between humans and technologies. This leaves the question of how visions of artificial agents reflect and affect the design and development of technology. As the next chapter will show, the metaphorical concepts that support these visions are interpreted in different ways for different reasons. I will turn my attention to the instrumental roles of metaphors in research and development practices in order to examine to what extent these visions can tell us something about how the envisioned technologies will relate to humans.

3. PERSPECTIVES ON COGNITIVE SYMBIOSIS

During the nascent years of the digital computer, Jospeh Licklider, widely regarded as an early pioneer of what we have come to know as the Internet, envisioned a new role for this machine. He advocated and actively pursued his vision of humans and computers tightly coupled in a "man-computer symbiosis". Through the metaphor of symbiosis, he sketched an image of future humans and computers as two "dissimilar organisms living together" tied to each other in an "intimate association" by their mutually benefiting interactions (Licklider, 1960, p. 4).

> The hope is that, in not too many years, human brains and computing machines will be coupled together very tightly, and that the resulting partnership will think as no human brain has ever thought and process data in a way not approached by the information-handling machines we know today.
>
> *(ibid.)*

Licklider's envisioned symbiotic relationship would be one of cooperation in "formulative thinking": computers would support humans in reasoning through technical problems by interacting with them in 'real-time' and in a more comprehensible way. He imagined that they would gather and transform data, recognize patterns, convert hypotheses into testable models, simulate mechanisms and models, carry out procedures, interpolate, extrapolate, and transform. Although formulating goals and hypotheses, taking initiative and evaluating the outcome would remain the province of humans, Licklider estimated that it would become increasingly difficult to "neatly" distinguish between the contributions of human operators and their equipment in the analysis of many operations (p. 7).

Licklider's ecological and cybernetic infused vision of future *cognitive* relationships between humans and technologies still resonates in agent discourse. The AmI vision, for instance, has a particular symbiotic flavor with its emphasis on intelligent interfaces that enable 'intuitive' and 'seamless' relationships, through their ability to 'anticipate' and 'adapt' to humans. The image of symbiotic relationships is also a recurrent element in agent research. It often appears in projects that focus on exploring the possibilities of building an artificial agent or a collection of agents that are part of an *adaptive system* encompassing both humans and technologies (Luck et al., 2005). The metaphor of adaptive systems links to a host

of additional metaphors, including self-organization and learning, which have served to distinguish artificial agents and multi-agent systems from conventional computer technology (Franklin & Graesser, 1997; Luck et al., 2005; Maes, 1994b; Nwana, 1996; Wooldridge & Jennings, 1995). In agent-based research, these concepts shape visions of computer systems that do not need be told beforehand exactly what the world looks like and what can be expected. At the same time they serve an instrumental purpose in structuring a way of thinking about the design and development of computer systems, as they provide conceptual tools to highlight and hide particular aspects of complex problems. In these different roles these metaphors do not necessarily have the same meaning.

Like Licklider's human-computer symbiosis metaphor, the abstract concept of adaptive agent-based systems has a suggestive element that opens the door to speculations about and promises of agents with enhanced abilities. In particular, it triggers a reconsideration of their ontological status in terms of *epistemic agency*. Licklider's vision of human-computer symbiosis harbors an interesting ambiguity. Although the notion of symbiosis implies two distinct entities, Licklider's vision derives its suggestive power from similarities. The characterization of humans and computers as two "dissimilar organisms living together" suggests that these two entities, although dissimilar, belong to the same class of 'living organisms' that can be described in terms of a set of general principles. As such, this vision leaves significant room for speculation on dissolving boundaries between the two entities in question. It facilitates the tendency to equate increased 'intimate coupling' between humans and technologies in an extended system with the notion that humans and technologies move progressively closer in terms of their ontological status. Similarly, the association of self-organization and adaptation with decidedly cognitive concepts such as learning, knowledge generation, and decision making, inspires some agent-enthusiasts to propose that increasingly self-organizing adaptive technologies will lead to relationships where it is no longer clear where the human begins and the agent ends (Clark, 2003).

This chapter explores the context-dependent nature of metaphorical concepts and the role they play in developing technology. As I argued in the previous chapter, a preoccupation with blurring boundaries can lead to a narrow, decontextualized view, which obscures the multiple interdependencies and constitutive asymmetries between humans and technologies. Moreover, as result of a seductive rhetoric the metaphorical character of concepts is easily overlooked. Consequently, narratives

about the promises of artificial agents provide a problematic basis on which to consider and evaluate technologies currently under development. The present chapter focuses on the context-dependent nature of the meaning of metaphorical concepts in agent discourse in order to further investigate the role of visions of artificial agents in research practices.

The first part of this chapter looks at what agent researchers do with metaphors, such as 'adaptive' and 'self-organization', and vice versa, what metaphors do with researchers. I will consider the role of these metaphors, and the multiple meanings they acquire, in the context of an illustrative example of a Dutch research project. I then turn to the *theory of Distributed Cognition* as introduced by Edwin Hutchins to offer an alternative perspective on adaptive symbiotic systems composed of humans and technologies, in which the asymmetries between humans and technologies play a key role. The discussion will show that the concepts that current research projects enlist cannot unproblematically be taken out of their contexts. In the final part of the chapter I reflect on the problematic conceptual aspects of abstract accounts of future adaptive agents.

3.1 ADAPTIVE SYSTEMS

When reading agent literature, we might be tempted to think that one significant problem of conventional computer systems is their limited ability to respond flexibly to new unanticipated events and changing requirements. Agent advocates typically characterize conventional computers as rather *rigid* devices in comparison to the idealized image of humans. We already saw some examples of this in the previous chapter. Unlike current day computers, so the argument goes, humans have the ability to respond to new, unanticipated events relatively easily. They can adjust their behavior and routines in response to changes in the environment without a complete *breakdown* of the process or action of which that behavior is a part. In a world filled with unknown variables and unanticipated events that intrude upon their habits and fixed beliefs, a certain level of *flexibility* of thought allows them to deal with this chaotic and messy world. In addition, humans learn new skills and new knowledge to deal with similar events more effectively in the future. Critics of the claims of early AI research have also argued that the human ability to adjust to changing circumstances and learn from them, individually or collectively, distinguishes humans from technologies. One feature that

distinguishes humans from machines is their ability to adjust to unanticipated events or changes in the environment that do not match predefined plans, rules and protocols (Suchman, 1987). To a certain extent, humans are able to modify and correct their actions, their knowledge or the environment in which they are operating, in order to amend their own and others' errors and mistakes (Collins & Kusch, 1998).

Conventional computers appear considerably limited by comparison. They are constrained in their scope of actions by their formal rule structures, specified by humans. They operate on rudimentarily stable facts and knowledge about human behavior and the real world, unable to modify this knowledge in response to external stimuli. Computer systems tend to work well when the problem is well-defined and when it operates in a *closed* environment. However, as soon as they are brought into *open* environments where problems can no longer be fully specified, it becomes increasingly difficult to develop a foolproof computer system that can appropriately respond to every event.[1] The rigidity and 'brittleness' of conventional computers have led to a wide range of research projects, aiming to amend this perceived shortcoming of computers.

The recurrent metaphor of computers as adaptive systems in computer science provides a way to structure, understand and talk about computers with some level of flexibility (Zambonelli and Parunak, 2002). It has acquired a number of meanings, in particular in various areas of research on intelligent technologies. Ask one agent researcher and they will refer you to the AI-inspired *machine learning* literature or game theory algorithms, ask another and they will start talking of biologically inspired *complex dynamic systems*. The term machine learning is traditionally associated with the logic-based symbolic algorithms that adjust the behavior of a system over time. It modifies its internal structure to *improve* or *optimize* the performance of the system on a particular task, given a set of training examples or in response to new events (Mitchell, 1997). Symbolic learning algorithms adjust domain knowledge or world models on the basis of examples and feedback functions. Domain knowledge can take the form of, for example, logical predicates or relational descriptions between facts. As an idea still closely associated

[1] In AI, a recurrent obstacle in the pursuit of intelligent machines is the *frame problem*. It expresses the difficulty of deciding what is relevant and what is not. When a robot takes an action in the world, like grabbing a cup, the world will change. But how does the robot update its knowledge about the new state of the world, without re-evaluating each single fact it knows about the world, such as the location of the cup or the color of the cup? For an overview of the frame problem as it is described in AI, philosophy and logic see Shanahan (2006).

with the more traditional concept of AI, the emphasis in these approaches is less on the interaction of the machine with its environment than on knowledge representation and (logical) 'reasoning' mechanisms, such as induction, inferences and pattern recognition.

Biologically or ecologically inspired research projects emphasize emergent behavior, often as result of interactions with the environment. An *Artificial Neural Network,* for example, 'learns' to relate input to output patterns by repeatedly adjusting the weights on the connections that link nodes within the network (Bishop, 1995).[2] Similarly, *reactive* and *evolutionary* methods place a strong emphasis on behavior *emerging* from local interactions between various simple components and the environment (Bonabeau et al., 1999; Brooks, 1991; Gershenson & Heylighen, 2003; Rocha, 2001). Rodney Brooks has been a leading figure in advocating a reactive approach to computing. His creations do not maintain an explicit representation of the world. Instead, they directly *react* to external events and adjust their internal structure in response to these events (Brooks, 1991). The 'adaptive behavior' of such a reactive system is different from machine learning in that it does not retain knowledge representations about previously encountered examples or events, and it is not explicitly aimed at improving performance. Evolutionary approaches, such as Genetic Algorithms, take their inspiration from Darwinian principles of evolution, such as survival of the fittest (Michalewicz, 1999). Such algorithms solve optimization problems by iteratively generating new 'populations' of solutions that are produced by evolutionary mechanisms, such as 'selection', 'mutation' and 'cross-fertilization'. More recently, Swarm Intelligence has been added to the list to describe the emergent properties of collections of agents (Bonabeau et al., 1999). This line of research is based on the idea that intelligent behavior can emerge from a collection of relatively simple entities interacting with and through the environment, analogous to natural systems, such as ant colonies. The distinctions between the different interpretations of 'adaptive systems' are not clear cut, and many research projects use some form of hybrid approach to develop adaptive technologies.

---

[2] Artificial Neural Networks are often referred to as *subsymbolic* methods, because they work "below the symbolic level". In particular, the level of computation is distinct from the level of representation. In symbolic methods a symbol or *computational token* represents a concept. A symbol is attributed meaning. In subsymbolic methods representations result from particular patterns of manipulation of computational tokens (Chalmers, 1992).

The discourse on artificial agents draws on and contributes to research on machine learning and adaptive systems. For example, agent researchers have explored various methods that allow 'intelligent assistants' to learn from user behavior or to respond appropriately to changing environments. In addition, researchers have contributed to existing theories and techniques through studying them from the perspective provided by the agent metaphor. In this chapter, I want to take a closer look at efforts to build adaptive or self-organizing multi-agent systems (MAS).

MAS research often builds on broader biological, systems theory and cybernetic discourses. These discourses convey a holistic world view that figures entities such as humans, biological organisms, financial markets, firms, as well as the Internet, as complex adaptive systems situated in open environments. These systems comprise heterogeneous interacting units, from which organization and complex behavior emerge. The common denominator in the various conceptualizations of such systems is the idea that they are in some way able to modify their behavior and *reorganize* or *reconfigure* their internal structure through time and in response to the changes in their environment. This world view underlies scientific research projects, theories and paradigms that aim to understand and describe the perceived complexity of, among others, organizations, the Internet, air traffic control, and warfare (Holland, 1996; Lewin, 1993). In particular, it supports conceptualizations of control processes, information distribution and decision making as situated in large scale and complex systems of humans and technologies.

A cybernetic or systems theory world view has provided heuristics for research on and development of MAS (Bullock & Cliff, 2004; Zambonelli et al., 2003). Thinking of computers systems in terms of complex self-organizing systems of agents offers an alternative to the more traditional models of computer systems that emphasize centralized, hierarchical control and stable structures. Luck et al., for instance, state that self-organizing agent-based technologies provide a "design metaphor" that encourages researchers to conceive of a computer system "as comprising interacting autonomous entities, each acting, learning or evolving separately in response to interactions in their local environments" (2005, p. 25). Agent researchers Zambonelli and Van Dyke Parunak claim that the idea of agent-based systems capable of 'adapting' and 'learning' provides new opportunities to model and build complex systems (2003). They argue that rather than specifying the behavior of a computer system at every level, embracing metaphors like adaptive

systems will encourage designers to think in terms of *manipulating* emergent system behavior. The suggested advantages of the MAS metaphor is that it provides a way to model complex systems in open environments, in which the heterogeneous components of the system are not known in advance and can change over time.

One specific example of a domain that, according to agent researchers, can be represented in terms of complex adaptive systems is that of disaster or crisis management (Kitano et al., 1999; Nathan Schurr et al., 2005). Coordination of a disaster response effort presents a problem in which various distributed heterogeneous components (e.g. emergency workers, various technological systems and material resources) need to be organized within a highly dynamic and time sensitive environment. The Combined Systems project that I will discuss in the following section provides an example of a proposed solution based on the metaphor of self-organizing MAS. I will take a closer look at this project to highlight different functions and meanings of the concept of adaptive or self-organizing systems in particular contexts.

An analysis of what agent researchers do with metaphors helps to get a better sense of the malleability and context-dependence of metaphorical concepts. The Combined Systems project provides a good starting point for this analysis, as it presents an ambitious and integrative approach to building agent-based computer systems structured around a vision reminiscent of Licklider's human-computer symbiosis. This vision guides and draws together a series of research projects that focus on different levels of system design. A full account of metaphors 'in action' would require a more extensive analysis of the multiple sites where metaphors are enlisted and used.[3] My aim here is to highlight some of the different ways that metaphors describe computer technology. I will, therefore, only concentrate on the concepts of adaptive or self-organizing systems as they appear in the descriptions of the central design vision and of one particular sub-project of the Combined Systems project.

3.2 VISIONS AND DESIGN METAPHORS

The *Combined Systems* project ran from 2002 until 2006 and was a collaborative effort of four Dutch research centers and three subcontractors supported by the Dutch Ministry of Economic Affairs (Burghardt,

[3] For an example of a more detailed account of the function of metaphors in system design see Mambrey & Tepper (1996).

2004; Storms, 2004b). The ambitions and motivations of the project are captured in the acronym *Combined,* which stands for **C**haotic **O**pen world **M**ulti-Agent **B**ased **I**ntelligent **NE**tworked **D**ecision support. The project proposed to develop a host of technological tools to support communication, coordination, collaborative decision-making and information sharing between multiple, heterogeneous parties in dynamically changing or chaotic environments. These tools were intended to support the realization of 'Combined' systems, defined as a network of heterogeneous distributed systems. The envisioned Combined systems would automatically configure at 'run-time', contributing to the ability of organizations to quickly adapt to unpredictable situations. Intelligent agents were viewed as the "enabling technologies to provide interoperability, communication, interaction and coordination" between the subsystems (Storms, 2004b, p. 139).

The project focused on crisis management as a target application area. A central assumption in the project was that in crisis management not all the variables and elements of the problem can be fully specified. To further explore this assumption and to put the different components of the project "into context", a hypothetical scenario was defined and presented as a "case study" (Storms, 2004b). This "validation scenario" describes a hypothetical crisis situation in the Rotterdam Harbor. Following the collision of two ships, a cloud of poisonous gas spreads across the city of Rotterdam, calling for a large-scale crisis response effort. The police, the fire brigade, hospitals and traffic control need to work together to evacuate buildings and streets, to provide medical care for the casualties, and to organize traffic in order to minimize immediate and long-term casualties and damages. The imagined crisis presented a scenario with many uncertainties, as the sequence and scale of events in such a crisis is difficult to predict or anticipate in advance. Many factors, including the weather, the number of people involved and the available resources, will determine the outcome of the situation. Further challenges that the Combined community envisioned included non-interoperable information systems, the exponential increase of information, contingent circumstances, and conflicting interests of the emergency services. The focus in the Combined Systems project was on problems that involved generating and maintaining overall 'situation awareness' and coordination in crisis response efforts.

As for the envisioned Combined technology, the project assumed that a fixed decision-support system structured on a plan-based hierarchical approach is ill-equipped to deal with the inherently unpredictable events

of crisis situations (Storms, 2004a). Such a system is vulnerable to breakdown as a result of damage. Moreover, a computer system in which lower-level execution components are governed through centralized higher-level control is impractical due to the possible exponential increase of information, and changing requirements. Instead, the project aimed to develop a range of technologies that would automatically assemble a decision-support system to meet the operational requirements at hand. Decentralized control in this system would support 'local interactions' between existing information systems and mobile devices in order to coordinate information, human actors and material resources. The system would be open, in the sense that "humans, sensors, actuators and other computational resources can join or leave the system" at any point (Storms, 2004, p. 139). In the event of an escalating crisis situation the system would be able to 'scale-up' by automatically incorporating new resources. In addition, the support for local interactions between components would enable the system to reconfigure in response to a loss of components. Thus, the size and shape of the system would vary in accordance with the changing requirements that a response effort generates.

A central feature of the Combined Systems project was the emphasis on an inclusive view of organizations that incorporates both humans and technologies. The project leader Paul Burghardt notes that during the project, considerable time was spent on studying documentation on escalating crisis situations to develop the *Combined Systems view* (Burghardt, 2004). This particular "conceptual point of view on Crisis Management Systems" presented future Combined Systems as collaborative, self-organizing networks of human actors and artificial agents operating in chaotic open environments (p. 52). It built on the vision of *Actor Agent Communities* (AAC), formulated by the Delft Cooperation on Intelligent Systems (D-CIS) lab.[4] Reminiscent of Licklider's man-computer symbiosis, the D-CIS research lab posits the AAC vision as solution to complex control problems:

> Many computer scientists, long forgotten and contemporary alike, have predicted a future in which humans and artificial systems work together in close fashion, even to the extent of being peers. And though progress on this matter has not advanced as quickly as some have predicted, there

[4] The D-CIS lab is a partnership between the University of Amsterdam, the Delft University of Technology and Thales Nederland. Thales Nederland is the Dutch division of the Thales Institute: a global organization focusing on electronics and systems and serving defense, aerospace and security markets (see http:\\www.decis.nl).

> is no denying that artificial systems have become an integral, elemental part of our world. Humans and machines are working together everywhere in contemporary society. The DECIS [*sic*] group emphasizes this relationship by viewing modern society as a collection of communities that consist of human actors and artificial agents: actor-agent communities (AACs).
>
> *(D-CIS website, retrieved May 20th 2007)*

Like constructivist studies of technology, the AAC vision underscores the role of both humans (actors) and technologies (agents) in control and decision-making problems. Communication, decision making, information sharing and coordination are conceived of as processes within a community of human actors and artificial agents. The D-CIS lab, however, presents the AAC vision as a "paradigm shift" in building and designing complex information systems.[5] It is a prescriptive design vision that specifies a number of desired qualities of these systems. In actor-agent communities, humans and software agents should "collaborate as peers" towards a "shared goal" or a "common mission" (Wijngaards et al., 2004). The Combined Systems view adopted this vision and presented future crisis management systems as hybrid networks of human actors and artificial agents making use of lower-level information systems. "The initiative in processes will alternately be taken by actors and agents, thus giving rise to mixed-initiative systems" (Brughardt, 2004, p. 53). The agents in this vision thus seem to be attributed a kind of agency that current technologies lack.

The Combined Systems view played a significant role in the project. It provided a design vision that simultaneously served to position the project in relation to other 'conventional' research projects as exploring an innovative integrated approach to developing technologies for crisis

[5] The D-CIS lab focuses particularly on the development of information and communication systems in the coordination of crisis response efforts. The AAC is elaborated and explored in two projects: the ICIS and Combined project. The **I**ntelligent **C**ollaborative **I**nformation **S**ystems project is still running. This project brings together industrial and academic partners within Consortium financed by the Dutch Government. The project, which commenced in 2006 and will run until 2009, aims to establish "a centre of excellence" that specializes in researching and developing "interactive collaborative information systems for the support of decision making in complex dynamic environments" (See ICIS website: http://icis.decis.nl). The ICIS project emphasizes the intelligence of artificial agents more than the DECIS project. However, I have chosen to concentrate on the Combined Project as it takes an integrated approach that more clearly illustrates the different levels at which metaphors are used.

management.[6] A key aim, as one internal report states, was to build a prototype system that could serve as a research platform to test and evaluate technologies and "new concepts" (Storms, 2004a). In particular, it adopted a more "organic" view that acknowledges human factors in organizations, in contrast to the "mechanical view of organizations" in the "traditional design" view of information systems (Burghardt, 2004, p. 55). The Combined Systems view emphasized decentralized organizations and the active role of agent-based technologies in these organizations as relatively new research areas.

Besides providing a 'new' design vision, the Combined Systems view also tied together the previously unrelated research projects of the project partners. The research conducted under the header of the Combined Systems project encompassed a diverse range of projects that addressed various aspects of system design and human/technology relationships. The promotional website repackages the results of these separate research projects as nine "building blocks" for Combined Systems.[7] The notions of self-organization and adaptivity supported these instrumental roles of the Combined Systems view, but to what extent do they apply to the technologies developed in the Combined Systems project?

So far, the envisioned artificial agents remain abstract elements in the design vision. In what sense would the Combined technologies contribute to the ability of Combined Systems to self-organize and scale up to dynamically changing environments? With the emphasis on research on multi-agent systems, the project committed to exploring the potential of self-organizing or adaptive qualities of these agents. Burghardt, at one point, suggests that MAS will only take care of the "well-structured

---

[6] See the Combined Systems promotional booklet *Combined Systems: Combining more for crisis management*. In this booklet the project is introduced as follows: "The Combined Systems project is one of the first integrated crisis management projects in the Netherlands. The project's contributions include: (1) a new model for the development of crisis management support systems: the Combined Systems view (2) new technology in the form of **intelligent building blocks** and (3) a diverse and dedicated **crisis management research community**". (http://combined.decis.nl/images/deliverables-/combined-project-booklet-2006.pdf).

[7] See http://combined.decis.nl/. Some examples of the nine building blocks listed on the website are: a software component that builds communication networks between mobile devices "on the fly", a *Semantic Network Engine* that implements a knowledge base in which electronic messages are stored, analyzed and redistributed; interface tools to support communication based on an *icon* language; an interface tool to support "critical thinking" about the developing situation; and coordination strategies based on ant-based routing algorithms.

tasks" (2004, p. 53). These well-structured tasks apparently include self-organization, for he proposes a few paragraphs later that the notion of *self-managing distributed systems* (SMDS) is an "important theme" in the project. This notion, he writes, can be applied "to low level technical configurations in information systems, but also to higher-level processes where multi-agent systems *and* people dynamically reorganize themselves as the situation in a crisis changes" (ibid., emphasis mine). The general idea behind this, he notes, is that technological systems automatically configure themselves to provide "certain services", such as automatically distributing information to multiple mobile devices or automating route planning to aid evacuations.

In another internal report *self-management* is explained in terms of *self-forming, self-organizing* as well as *self-healing* (Storms, 2004a). A self-managing system, the report stresses, will have to be capable of regulating itself in terms of system integrity and functional behavior. It has to integrate distinct systems, to organize the distribution of information across the different systems, and "to detect and remedy anomalies in the planned execution of tasks" (p .4). These different interpretations of self-management emphasize the ways in which Combined Systems are envisioned to change their internal structure in response to or irrespective of changes in the environment. The self-management spoken of in the Combined System project thus suggests a level of autonomy of the system, in the sense that it can operate for a period of time without direct control. I shall return to autonomy in the next chapter. What is of interest for the current discussion is the suggestion that the MAS technologies developed in this project would be capable of operating (more) flexibly in an open complex and dynamic environment as a result of their adaptive or self-organizing quality.[8]

What exactly does it mean for a multi-agent system to be capable of self-organization? And how do agents feature in this? Burghardt notes that:

> The point we have come to realize at Decis [*sic*] in general and in the Combined Systems Project, in particular, is that the qualities of systems

[8] The difference between adaptive and self-organization is not well defined in the field of agents research. Not surprisingly, as they are metaphorical concepts used to describe some salient features of computer technologies in different contexts. However, sometimes a distinction is made between adaptive and self-organizing on the basis of the distinction between the behavior of a single agent or of a collection of agents. Self-organization can be an emergent property of a collection of relatively simple agents. An adaptive agent, then, implies a more complex entity (Luck et al., 2003).

> for collaborative decision-making are not only qualities of technical systems or qualities of human networks, but qualities of the complete configuration of human and artificial systems. The consequence of this insight is that quality terms such as adaptability, flexibility, scalability, trustworthiness, efficiency, etcetera take on a different meaning. [. . .] The quality of a Combined Crisis management system should not be confused with the quality of its information system or of its human organization.
>
> *(Burghardt, 2004, p.55)*

Burghardt distinguishes between different kinds of adaptability and flexibility. His comment underlines the ambiguity in the concepts enlisted to characterize Combined Systems. The self-organizing quality of the envisioned systems as a whole - that is those systems constituted by humans and technologies - is different from that of the multi-agent systems. To get a sense of the meaning of the envisioned self-organizing quality of multi-agent systems in the Combined project, I will take a closer look at one particular 'building block' of the project.

3.3 DISTRIBUTED PERCEPTION NETWORKS

As part of the Combined project, a team of researchers at the University of Amsterdam (UvA) focused on the development of a system that could support human operators in handling the large amounts of heterogeneous sensory data and information needed in decision-making processes (Maris & Pavlin, 2006; Pavlin et al., 2004). This effort centered on the idea that in crisis-response efforts human decision makers have to be capable of rapidly assessing the situation, based on a vast number of different information flows. Decision makers are confronted with large amounts heterogeneous noisy data and information from which they have to infer and extract relevant information. In one paper that describes the team's research project, Maris and Pavlin note that new communication and sensing technologies, including GSM devices, cameras or gas detection devices, have contributed to a growing amount of valuable information (Maris & Pavlin, 2006). It has become increasingly difficult to manually process this body of information. They proposed that "situation assessment in a crisis situation can be improved by technological support systems" that take over part of the gathering, interpreting and distribution of information (p. 377). The contribution of the UvA researchers to the Combined project, therefore, entailed a research project in which they developed and experimented with what they called an "automated information fusion system" that 'fuses' data

and information from various spatially dispersed heterogeneous sources and presents only relevant information to human operators.

Information fusion for crisis management, Maris and Pavlin point out, is an emerging field of research in which the focus is on formulating methods and algorithms to automatically infer "hidden" events from data and information, based on knowledge of the domain (p. 376). More specifically, the researchers conceived of fusion as "a mapping from different observations to potential causes, i.e. backward reasoning from symptoms to their causes" (Pavlin et al, 2004, p. 466). Information fusion, according to researchers, entails estimating the *likelihood* of events given some observable data.

The various papers describing the efforts of the team stressed that in crisis management, a centralized hierarchical approach to information fusion would be vulnerable to a single-point failure, and would "suffer from inadequate communication and processing capacity" (Pavlin et al., 2004, p. 466). Moreover, they assumed that in dynamic situations (e.g. where the configuration of the system can change during the operation) centralized control is computationally costly, as continuous centralized reasoning is required about the states of the fusion system to ensure a valid fusion process. This can result in an exponential growth of information and additional processing. The team, therefore, proposed an alternative approach to information fusion based on the idea of self-organizing and adaptive multi-agent systems.

The researchers defined their approach in terms of what they called *Distributed Perception Networks* (DPN). In their words, "a DPN is essentially an organization of agents which maps large quantities of evidence to the hypotheses of interest through cooperation" (Pavlin et al., 2005, p. 802). The envisioned networks were defined as a multi-agent system that implements a software layer on top of "existing sensory, communication and processing/storage infrastructure" (Maris & Pavlin, 2006, p. 376). Defined as such, interacting agents enable the system to automatically assemble 'ad-hoc' networks of mobile devices and dedicated sensor technologies. Within these networks agents gather information about the area and infer the likelihood of particular events, such as the presence of fire at a certain location. Some of the agents perform fusion tasks. Based on input obtained from other agents, these *fusion agents* use 'local world models' to derive higher-level information in the form of the probability of a particular event. Other agents represent either a mobile device operated by a human or a computer controlled sensing device. These *sensor agents* serve as a 'wrap' around humans or sensory devices to allow

fusion agents to communicate with these information sources. This decentralized fusion system, according to the researchers, would "allow for robust and adaptive fusion" (Pavlin et al, 2004, p. 466).

Maris and Pavlin provided an example, based on the Rotterdam harbor scenario, of the type of fusion problem that the DPN system should be able to cope with (Maris & Pavlin, 2006). They imagined ammonia escaping from the collided ships, forming a toxic cloud over a densely populated neighborhood. To assess the level of threat and decide upon actions, such as a possible evacuation, decision-makers must quickly determine the type of gas and its concentration. Inspired by the pervasive mobile communication devices and infrastructure already available in real-world situations, Maris and Pavlin envisioned a DPN that gathers information from various sources present at the scene to generate and test a number of hypotheses about the type and concentration of gas. They identified mobile phones, personal digital assistants (PDAs) as well as dedicated sensors as potential sources that could provide observations of varying quality about symptoms of the presence of ammonia. Prompted by an initial report of a single gas sensor about the presence of an unusual concentration of gas, the DPN software would evaluate several hypotheses about different types of gas by contacting and querying other sources. The agents in the system would automatically construct and configure several DPN networks, each of which would be geared to test for a particular hypothesis. If a network registers a sufficiently high probability for the presence of a particular type of gas, the DPN would alarm the operators in the control room.

A more detailed look at the specification of the system reveals that the notion of self-organizing multi-agent systems is a convenient shorthand to describe an algorithm based on *Bayesian Networks*. A Bayesian Network (BN) represents the probabilistic dependencies between a set of variables as a graph or network of connected nodes.[9] Nodes represent variables whose states correspond to events in the world. The arcs that connect nodes represent conditional probabilities between different states of variables. BNs can be used to calculate the probability distribution over events, such as fire, given the absence or presence of other events, like smoke or intense heat, assuming that the probabilistic dependencies between these events are known beforehand. The probability of a certain event is sometimes called a *belief*.

[9] For a detailed account of Bayesian Networks see (Jensen, 2001). A shorter but helpful explanation is provided in (Rich & Knight, 1991).

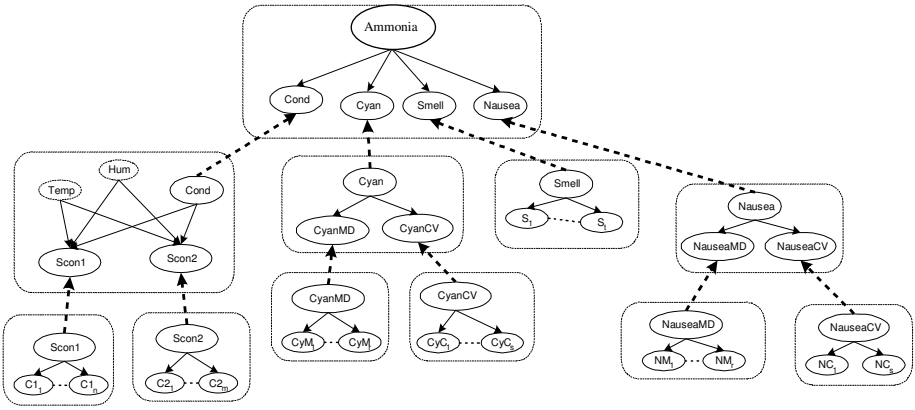

Figure 3.1. A DPN fusion organization where each dotted rectangle represents a DPN agent. Thick dashed lines represent communication between cooperating agents, which share partial fusion results. Each agent makes use of a local Bayesian Network that captures specific expertise over a certain domain, in order to evaluate the hypothesis "Ammonia" (reprinted from Maris & Pavlin, 2006).

Represented as a Bayesian network, the DPN researchers define a fusion problem as composed of a hierarchically ordered set of simpler problems. An agent in the DPN corresponds to a BN that represents a particular causal model or a "local world model' of a subset of relevant events in the world (Pavlin et al., 2005, p. 802). Maris and Pavlin provide an illustration of how a BN represents the fusion problem of estimating the probability of the presence of gaseous *Ammonia* in an area (see figure 3.1). This BN consists of a number of smaller BNs (agents). Each of the BNs has one 'root node' connected through a number of 'arcs' to a set of 'leaf nodes', in such a way that no cyclical loops are formed. Agents are connected through the 'root nodes' of the BN and their 'leaf nodes'. The fusion agents fuse information by updating the probability of the particular event that they represent, as new evidence is 'propagated' throughout the network.

The advantage of DPNs over other approaches to information fusion is, so the researchers claimed, that a DPN assembles itself at 'run-time', i.e. the actual scale of the network and the availability of information sources do not have to be known beforehand. Furthermore, DPNs allow for "distributed asynchronous propagation" (de Oude et al., 2005). Propagation of 'beliefs' along the branches occurs in a bottom-up fashion when new evidence becomes available. In other words, when a sensor agent (human or non-human) produces a new read-out, it triggers a 'belief' updating processes throughout the network. All the conditional probabilities that are affected by this evidence are re-computed. This

process allows for a varying number of sensing agents. Thus, at any one moment new agents can be added or removed. Finally, the architecture of the DPNs make it possible for fusion agents to actively search for other agents to 'request information' needed to update their 'belief' about the state of the world. The agents 'find' each other through what is aptly labeled a "yellow pages" agent, where agents "register" their "services" (i.e. what kind of information they have to offer to other agents) (Pavlin et al., 2005, p. 803). Agents can consult this particular agent to locate other agents that can provide the information required to update their world model. As a result no centralized fusion control or synchronization is required.[10]

In the detailed algorithmic specification of the DPN system the metaphorical character of concepts, such as self-organization and adaptivity, is apparent. The researchers used the concept of self-organization to partially explain the behavior of the system without having to fall back on the intricate mathematical equations and formal definitions. These metaphorical concepts act as, what agent researchers Wooldrige and Jennings refer to as *abstraction tools*. Abstraction tools are concepts that reference human properties or intentional notions and "provide us with a convenient and familiar way of describing, explaining, and predicting the behavior of complex systems" (Wooldridge & Jennings, 1995, p.119). They invoke the notion of *intentional stance*, as introduced by the philosopher Daniel Dennett, to explain how these abstraction tools work.[11] Artificial agents can be attributed beliefs, desires and so on, when it helps to understand the behavior and the structure of the system. An intentional stance to the explanation of a system is required when the *design stance* - i.e. that stance that will lead to a mechanistic understanding of the system - is not possible or practical due to the

[10] Distributed, decentralized systems, such as multi-agent systems, are often advocated on the basis of their flexibility, relative robustness and ability to deal with heterogeneous data. However, they introduce their own set of problems. For example, yellow page agents as a solution to the problem of connecting agents can be potential bottlenecks in distributed systems (Decker et al., 1997).

[11] Daniel Dennett describes an *intentional stance* as "the strategy of interpreting the behavior of an entity by treating it as if were a rational agent who governed its 'choice' of 'action' by a 'consideration' of its 'beliefs' and 'desires'" (1996, p. 27). This strategy can be usefully applied when it allows an observer to predict and thereby explain the actions of a complex entity. Dennett contrasts this strategy with the *physical stance* and the *design stance*. The former is the method that explains entities whether designed, alive or not alive, in terms of the laws of physics. The latter is the strategy of explaining or predicting the behavior of an entity based on assumptions about how its design is supposed to operate. For more details see (Dennett, 1993, 1996).

complexity of the system. In the same way, concepts referring to natural phenomena can be employed as abstraction tools. Thus, a DPN can be more easily explained to a wider audience, by describing it in terms of agents that, through local interactions, configure, organize, and assemble themselves into an ad-hoc network, instead of by detailing the formal rules and mathematical intricacies of the algorithm.

Abstraction tools not only provide a convenient way of explaining system behavior, they also bring to the fore particular aspects of the world that other abstraction tools would hide, providing alternative ways of thinking about a problem. Thus, the idea of a self-organizing collection of agents allowed the DPN researchers to restate the problem of distributed information fusion in complex, chaotic environments into the more manageable problem of coordinating a number of discrete agents. This conceptualization shifts the problem from designers anticipating and planning all possible sequences of events for the entire system, to defining different components and their possible local interactions. Moreover, through this conceptualization the DPN researchers positioned and distinguished their approach with respect to other approaches to dealing with information fusion. The characterization of the DPN as 'adaptive' points to the distinguishing, more generic methods the researchers employ with the aim of tackling a wider class of information fusion problems.

Although the researchers do not explicitly define the concepts 'self-organization' and 'adaptive', they acquire a particular meaning in the DPN approach. As an abstraction tool, these concepts apply to a computer system that is defined within a narrowly defined theoretical and hypothetical 'world view', which isolates and stabilizes particular elements of the 'real world' and excludes others. The various papers describing the DPN application and the underlying algorithms reveal a number of abstraction steps that allowed the designers to focus on the algorithms and to formulate 'well-structured' tasks. For example, information about natural phenomena was reduced to a probability function based on yes or no events. Humans were represented as agents that have the sole purpose of providing information. In addition, the DPN approach assumed a finite set of predefined modeling *parameters*, i.e. the BNs only capture the conditional probabilities of a known set of variables. The implemented computer system is, therefore, adaptive only to the extent that it can deal with varying numbers of agents and uncertain information about known concepts. Adding a new type of variable or concept - such as new types of evidence for the presence of

ammonia or even an unanticipated kind of toxic gas - would require human intervention. Similarly, a limited range of valid network configurations, predefined by the designers, constrains the level of self-organization of BN-based systems. The proposed architecture requires the full-specification of a probabilistic model that describes the possible causal independencies between the nodes in the Bayesian Network. In one paper, the DPN researchers mention that "in general the parameters are found through domain experts, automated learning methods, or a combination of both" (2006, p. 727) This role of humans, however, is left unexplored. The various papers describing the approach mention the work of human designers, operators or decision-makers only in passing. A system conceived of as encompassing the DPN technology and its human designers and/or operators in this respect is adaptive on quite a different level, for such a system is capable of adjusting to new variables.

The example of the Combined Systems projects demonstrates that adaptive or self-organizing systems are useful metaphorical concepts within specific contexts. The danger of extracting these concepts from their contexts is that a more common or different usage connotes qualities of humans or natural systems, which are filtered out in the more specialized usage of the abstraction tools in development practices. The suggestion that a DPN system is adaptive, without further qualification, might lead to mistaken beliefs about the abilities of the system, based on a comparison with human behavior. Moreover, the representations of adaptive systems in the DPN approach and the Combined Systems view build on abstractions of human/technology relationships that filter out the complexities of the interdependencies between humans and technologies. When considered in isolation from their context such conceptions of 'adaptive systems' can lead to the idea that properties of humans and technologies can be unproblematically transferred from one entity to the next. The next sections will discuss an alternative interpretation of 'self-organizing hybrid systems' which challenges this idea.

3.4 DISTRIBUTED COGNITION

The theory of Distributed Cognition (DC), as developed by Edwin Hutchins and his colleagues (Hollan et al., 2000; Hutchins, 1995; Hutchins & Klausen, 1996) offers a detailed view of hybrid systems that acknowledges the particularities of humans and technologies and their interdependencies in cognitive processes. DC embraces the idea of viewing cognitive processes as extending beyond the human brain and

distributed over internal (the brain) and external (the social and material environment) structures. I will focus, in particular, on Hutchins' elaboration of the theory. Hutchins argues for a shift in focus from the individual, information-processing brain as the sole locus of cognitive processes, such as those involved in memory, learning, reasoning, and decision making, to a more inclusive view. He pursues a "softening" of boundaries in "social space, in physical space and in time" that have been established "primarily for analytical convenience" by previous cognitive science approaches (1995, p. xiii).

DC is part of a trend in cognitive science to extend the focus of analysis for studying cognitive processes. Increasingly cognition-oriented theories have turned away from the dominant cognitivist tradition of the late 20th century that unified AI and cognitive psychology. In this tradition, theories of cognition were mostly concerned with what went on inside the human mind in reaction to certain stimuli. In response to critiques and the limitations of these models, cognitive scientists increasingly came to adopt and advocate more inclusive views of cognition, which acknowledge its (socially) "situated", "embodied" and "distributed" aspects (Clancey, 1997; Suchman, 1987; Winograd & Flores, 1987). Such approaches emphasize the significance of the body as well as of the social and physical environment in cognition. Moreover, technological artifacts have been recognized as essential elements in cognition. We are able to reason the way we do, because we employ the tools we make and use to manipulate our surroundings. For instance, most of us are unable to solve complex equations without using a pen and notepad. Tools are essential elements in structuring and organizing our ideas and thoughts.

DC, as defined by Hutchins, identifies three kinds of distribution within a cognitive system.[12] First of all, cognition is seen as a collective process emerging from social groups, differing from the cognitive processes inside an individual's brain. The interactions between members

[12] According to the theory of DC, cognition does not emerge from one type of cognitive system, but from various interconnected and often subsuming cognitive systems with different cognitive properties. This view of cognition implies a seamless (i.e. never ending) cognitive system. The theory therefore proposes a flexible delimiting of the unit of analysis for studying these processes "wherever they may occur". This should be determined by the "functional relationships" between the elements that participate in these processes, rather than their "spatial collocation" (Hollan et al., 2000, p. 175). In other words, the cognitive processes studied are not just those between elements that are physically connected, but also include those processes that extend over time and space. Functional is interpreted as operational dependence.

of a group, verbal and non-verbal, determined by the social organization among the members and the context in which the activity takes place, provide a medium through which knowledge and information can flow and by which they are shaped. Secondly, the human elements in these distributed systems are individually and collectively connected with and through the material world. "Humans create their cognitive powers, by creating the environments in which they exercise those powers" (Hucthins, 1995, p. xvi). In other words, cognitive processes extend beyond the individual brain, as humans recruit and exploit their physical environment in their reasoning processes, enabling them to perform cognitive activities that they would otherwise not be capable of. Finally, DC stresses that cognition is formed by a continuous process, building on accumulated knowledge that is crystallized and saved in social organization and material and conceptual technology. It is this process that Hutchins defines as 'culture' (Hutchins, 1995, p. 353). He perceives culture as a process that accrues "partial solutions to frequently encountered problems" (p. 353) and the "residua" of this process are tools, concepts and social rules. In DC, the three kinds of distributions are interdependent and analysis of one cannot be separated from the others. Nevertheless, the cognitive processes in a sub-system, such as the processes in the human brain, can be "radically different" from the processes in another subsystem, such as the system composed of a person in interaction with a tool or a group of individuals.

Hutchins adopts the traditional metaphor of cognitive science – cognition as computation – to describe processes in extended cognitive systems. These processes can be described in term of "computation realized by the creation, transformation, and propagation of representational states" (Hutchins, 1995, p. 49).[13] He defines computation in a broad sense as "the propagation of *representational state* across *representational media*" (p. 118). A representational medium can be a wide range of structures, such as the neural structures making up internal memory, the linguistic constructs of spoken language, gestures in non-verbal communication or the structure of physical and conceptual technologies. The representational state of a medium is "a configuration of the elements of a medium that can be interpreted as a representation of something" (p.

[13] By offering a (re)conceptualization of cognition as a distributed computational process extending beyond the individual mind, Hutchins wants to draw the focus back onto the social, cultural, historical and material dimensions of cognition. Hence his proposal that cognition should be studied outside of the laboratory "in the wild", because this "may reveal a different sort of task world that permits a different conception of what people do with their minds" (p. 371).

117). For example, consider how a mathematical problem is represented in various media during the activity of a mathematics teacher explaining a mathematical operation to a student. Knowledge about numbers and algebraic rules are represented by some neural patterns in a mathematics teacher's brain, but also by the words that the teacher utters to the student and their sequential order, by the configuration of the symbols written on the blackboard, by the act of writing the symbols on the blackboard, and by the neural patterns of the student that perceives these actions. The brains, the vocal communication, the blackboard as well as the act of writing are all representational media. Each medium represents the problem in a different way.

The processes within a cognitive system propagate representational state across representational media by "bringing the states of the media into coordination with one another" (Hutchins, 1995, p. 116). As information moves from one representational medium to the next its representation is transformed. This transformation is the result of the propagation mechanisms that adjust the information-bearing structures to coincide with each other. Hutchins turns to AI pioneer Hebert Simon to provide an illustration of how representational states can be propagated. Simon used *theorem proving* as a case to describe how problems are solved through a series of simplification steps that re-represent the problem until the solution becomes transparent. In his example of 'theorem proving', the computational system consists of a set of axiomatic propositions and a set of rules to operate on these propositions. The application of the rules, according to Hutchins, is the "means of coordination" between the rule and the state to which it is applied (p. 117). This kind of symbol processing through rule application is an instance of a broader class of computations. Some implementations of computation, Hutchins claims, cannot be adequately described in terms of symbol processing. The embodied perceptual actions involved in writing down numbers on a piece of paper in a particular configuration are an essential part of the act of solving a complex equation. These actions cannot be described in terms of symbols and rules without changing the nature of the cognitive process. Each element in the cognitive system transforms the problem, making different demands on *cognitive abilities*.

The different demands on cognitive abilities that different representational media make are a key point in DC. According to Hutchins, each medium has "physical properties that determine the availability of representations through space and time and constrain the sorts of

cognitive processes required to propagate representational state into or out of that medium" (Hutchins & Klausen, 1996). Thus, speech has the property of being ephemeral and linear, in contrast to the more durable image of a photograph which can convey more information at one time.[14] Replacing one medium for another thus calls for different cognitive abilities to perform the same task. For example, Hutchins and Klausen describe how the duplicate flight instruments in civil transport aircrafts provide "a redundant distribution of access to information that supports mutual monitoring between the crew members and is essential in the maintenance of intersubjectively shared understanding of and expectations about the situation of the aircraft" (2000, p. 13). As both pilots can see what the other is doing they can be expected to maintain a shared understanding of the situation, as they can deduce the intentions and consequences of actions. Restricting the redundant access to information by, for instance, building two separate work stations will change the coordination of representational state propagation. The pilots will have to explicitly communicate their actions to the other pilot, making it more difficult to create a shared understanding. Substituting a component in a cognitive system thus leads to a reconfiguration of the system and to a redistribution of cognitive abilities throughout the system.

DC sketches the image of cognitive systems as layered, adaptive systems, consisting of heterogeneous elements, interlinked by their interactions in which knowledge is generated and stored. The adaptive character of these systems is an effect of the interactions between distinct media that extend beyond a single time and place. The systems adapt over time as representational media are brought into coordination with each other. Hutchins defines learning as "adaptive reorganization in a complex system" (1995, p. 289). A cognitive system discovers and saves solutions to frequently encountered problems by reconfiguring the components in the system. By defining learning in this way, he aims to emphasize that, like other cognitive processes, an understanding of learning requires the recognition of the role of the sociocultural and material environment. Learning is the result of the interactions between media both inside and outside the individual, rather than something that only happens below the skin. This, however, does not mean that the cognitive processes involved in learning within an individual are the same as the processes between a number of humans or between the individual

[14] The properties of the representational media are similar to what Donald Norman calls *affordances,* discussed in Chapter 2. (Norman, 1999).

and an artifact. Individual learning is different from organizational learning (p. 349).

For Hutchins, the tools and practices humans employ in their problem-solving tasks are the residue of cultural learning processes. Technological artifacts embody formalized knowledge, in the form of rules, models, values, strategies, and heuristics. He calls this formalized knowledge *crystallized* knowledge. For instance, the "crystallized" knowledge in a nautical chart used for navigation on board a navy ship represents the accumulated knowledge of generations, specifying procedures, rules of computation and information about the world that "no navigator has ever had, nor will one ever have" (p. 111). The naval instruments that Hutchins describes have "internalized" the procedures for measuring something about the world. They capture and represent regularities in the world, which become *useful* when the device is *properly* manipulated. In coordination with the development of other artifacts, they are embedded in a network of "mutual computational and representational dependencies" (p. 114).

The theory of DC provides an analytical framework to study how humans perform cognitive tasks in interaction with technologies and their environment. The analytical framework links together humans and technologies on a cognitive level through the metaphor of computation. The focus is explicitly turned away from the internal processes in the individual brain, as well as more implicitly from other features of human behavior like emotion and moral responsibility. Although this provides a restricted and particularly functional view of the role of both humans and technologies, it highlights aspects of what we generally understand to be cognitive processes (e.g. learning, remembering, and problem solving) that have been overlooked within earlier cognitivist theories. By (re)conceptualizing cognition as a computational process extending beyond the individual mind, Hutchins explicitly aims to bring back into focus the social, cultural, historical and material dimensions of cognition. Although he wishes to "soften" the sharp boundaries drawn by the inwardly oriented cognitivist theories, the differences and asymmetries between humans and technologies do matter. A DC perspective, as a result, allows for a description of systems of humans and technologies that acknowledges the complementary and multiple functional roles of technological artifacts in transforming and simplifying problems. As such, it sheds a different light on the hybrid self-organizing systems envisioned by the Combined System view.

Similar to the Combined System view, Hutchins considers adaptivity to be a property of an extended hybrid system, but he emphasizes that the adaptive character of the system cannot be reduced to a particular quality of a system component. It cannot be described solely in terms of some skin-bound processes or some special properties of technology. Rather, the adaptive character of an extended cognitive system is the result of interactions between the various distinct media. From a DC perspective, therefore, the description of a DPN-based decision-support system as an adaptive cognitive system is incomplete if it leaves out the human actors involved. Moreover, the cognitive processes in this system cannot be explained without taking the sociocultural and historical context in consideration. The researchers, their programming tools, the hardware, their notes, their reference materials, the algorithms and all the other 'props and aids' that are involved in the process of developing computer systems constitute one cognitive system. Even the concepts of adaptive and self-organizing systems as abstraction tools serve a functional role in these cognitive processes, by providing the vocabulary that supports an understanding of the complex processes to be modeled. For the DPN researchers, the computer technology under development is not a tool to solve a problem in the external world, rather the system itself is a problem to be solved. In the process of solving this problem, they draw on crystallized knowledge, such as probabilistic models of real world events, to further stabilize and formalize knowledge about information fusion mechanisms. The implemented DPN algorithm is a representation of this knowledge. It is an abstraction of, in the words of Hutchins, "the operation of a sociocultural system from which the human actor has been removed" (1995, p. 363).[15]

[15] Hutchins argued that automations of "human" tasks rarely formalize what is in the mind of individual. Rather, they model the cognitive properties of a distributed cognitive system. Hutchins claims that early AI efforts mistakenly placed symbols in the mind, because "the computer was never a model of the person to begin with" (1995, p. 365). Instead, he says that "[T]he computer was made in the image of the formal manipulations of abstract symbols" (p. 363). He offers a perspective on the famous Chinese Room problem offered by Searle. In this thought experiment, a person inside a room exchanges received Chinese symbols for other Chinese symbols based on a set of rules. Searle argues that neither the person in the room nor anything in the room can be said to understand Chinese. Key to Hutchins' argument is that the sociocultural system that is the ensemble of the room, the person, the symbols and the rules seem to speak Chinese. This cognitive system has different properties than the sub-system that is the person's brain (Hutchins, 1995, p. 362).

Adaptivity as a feature of a distributed cognitive systems provides a different perspective on how decision-support systems can contribute to the self-organizing capacity of cognitive systems in the context of crisis-management. In the context of a crisis response effort, a decision-support system is part of a cognitive system different from the one that includes DPN researchers. To properly understand the role of decision-support systems in distributed cognitive processes we would have to perform an extensive analysis of cognitive systems "in the wild", to use another phrase of Hutchins. The same holds for understanding how a self-organizing support system would change these processes. Nevertheless, here I will draw on Human Factors research on 'situation awareness' to provide a background against which to consider some aspects of the role of decision-support systems in distributed cognitive processes.

The idea of *situation awareness* is a reappearing theme in the Combined Systems project.[16] This notion comes from the Human Factors researcher Mica Endsley. She defined situation awareness as the "perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the future" (Endsley, 1996, p. 164). This understanding of "what is going on" is a state of awareness of humans in complex processes. Recall that an important assumption of the Combined System project was that maintaining situation awareness in a crisis management effort becomes increasingly difficult for humans to accomplish as a result of growing amounts of information. The proposed solution to this problem, in particular in the DPN project, was to create tools that allow the coordinators to delegate certain information-processing and knowledge-management tasks to computer technology, by automating part of the information gathering, synthesizing, and distribution processes.

The implementation of the DPN relates to the DPN researchers in a distinct way, as compared to how an eventual DPN-based decision-support system would relate to human coordinators in crisis management situations. Computer systems can and have been enlisted in tasks as simulation devices, which model real-world phenomena (Luck et al., 2005). Human coordinators can use simulation devices to serve a similar

[16] In the Combined project the notion of *organization awareness* is also used to describe the type of understanding that a Combined System should enable. Organization awareness includes "an understanding of the multiple parties that make up the organization and how they relate to each other" on top of situation awareness (Oomes, 2004).

function to the role that computers fulfill in relation to DPN researchers, viz. as a tool that supports reasoning about phenomena. Adopting an 'intentional stance' towards these systems, i.e. conceiving of them as exhibiting adaptive behavior in the same way humans would, could potentially facilitate this reasoning. The DPN system, however, is not intended to serve as a simulation device. The task that the human operator supported by the system should perform is to continuously find an optimal solution for the coordination of the human and technical resources, given the existing constraints and a continuous stream of contextual information. In the context of crisis management the system would relate to its human operators in a more *hermeneutic* form, mediating the access of humans to the external world (Ihde, 2003; Verbeek, 2000). It provides a representation of the world that requires further interpretation of the human operator. Offloading information filtering or fusion to a system that is capable of adapting to changing circumstances can relieve the operator from certain tasks. However, it is not a sufficient condition for the enhancement of performance.

Human Factors research in aviation and military technology has long recognized the problem of automating cognitive tasks. Endsley's 'situation awareness' has contributed to a further understanding of this problem. Maintaining a high level of situation awareness requires *active* involvement of the human coordinator, as Endsley points out (1996). Good decision-making based on high-level situation awareness involves more than passive monitoring for the human operator. She argues that automating cognitive tasks can result in the 'out-of-the-loop' problem. Through empirical studies she and other Human Factors researchers have shown that situation awareness of human operators can deteriorate when they are assigned the role of observer or monitor (Endsley, 2001; Wiener, 1985). This deterioration can occur when humans rely too much on the accuracy of the automation, when they distrust the automation as result of a high rate of false alarms, when they do not have access to the feedback needed to support situation awareness, or when they are passive observers of the system rather than active processors (Cummings, 2004; Parasuraman & Riley, 1997). An individual who is not actively involved in the processing of information more easily loses track of what is going on and therefore it becomes more difficult for her to make appropriate decisions. In addition, in order to be actively involved the human operator has to maintain an accurate understanding of how the system works in order to effectively integrate the technology in her

reasoning processes. A more complex system can increase the cognitive load of the human operator and thus complicate the problem.

A DC perspective highlights the value of technology as a predictable structural medium in generating and maintaining situation awareness. Hutchins explains that what technologies contribute to the process of problem solving is that they reorganize the various tasks to be performed, in such a way that the cognitive tasks of the human are simplified. Useful technologies allow "people using them to do the things that need to be done while doing the kinds of things the people are good at: recognizing patterns, modeling simple dynamics of the world, and manipulating objects in the environment" (p. 155). Technologies take over the computations or algorithmic steps that distract from the actual task to be solved and set constraints on what can and cannot be performed. From this point of view, the behavior of a decision-support system simplifies the problem for the coordinator, when it embodies crystallized knowledge. The predictable transformative capacity of the system contributes to timely and appropriate decision-making. The idea of an adaptive or self-organizing computer system that takes over cognitive tasks, such as information gathering and filtering, is a problematic notion with regards to maintaining situation awareness. A 'self-organizing' system can make it more difficult for human operators to understand and anticipate the behavior of the computer system. An insufficient understanding can result in "automation surprises" with potentially disastrous effects (Sarter et al., 1997). A decision-support system that adapts in ways not authorized or recognized as valid changes can increase the cognitive workload of the coordinator and complicate her decision-making tasks. In the coordination of crisis management situations, unexpected system behavior that does not fit in with the expectations of the human coordinator can even be dangerous. In other words, the notion of self-organizing or adaptive technologies for information filtering can generate less adaptive distributed cognitive systems in the context of crisis management.

DC offers a view in which the asymmetries between humans and technologies are a constructive feature in human/technology relationships, rather than a problem to be overcome. Crisis management conceived of in terms of distributed cognitive processes demonstrates that a decontextualized notion of adaptive systems, which does not differentiate between the varying roles of humans and technologies, runs the risk of inhibiting and frustrating design processes. The suggestion that computer systems can be made to gather, interpret, filter and act on

information autonomously in complex, dynamic and chaotic environments, independent of human coordination ignores the problems of automation that have a longer history and that continue to challenge designers of information technology and automated systems. Such a suggestion overlooks the interdependencies between the roles of humans and technologies in sociotechnical organizations. The human operators and their skills and responsibilities need to be taken into account, when considering how the entire system should behave. From the point of view of DC, leveling humans and technologies in terms of their competences is not necessarily desirable in every context. Nevertheless, DC leaves open the possibility that under some circumstances the ambition to simulate human behavior is less problematic or even preferable.

Like the Combined System view and the DPN approach, DC builds on metaphorical concepts, such as cognition, computation and adaptive systems. Hutchins takes this cluster of associated concepts 'out-of-context', i.e. out of the context of conventional theories, to redefine them in a new domain. Conceiving of humans and technologies as interlinked in one cognitive, symbiotic system provides a powerful framework to conceptualize and study cognition, action and human/technology relationships. Such a framework poses the question what its metaphorical concepts hide and how these concepts relate to other conceptual frameworks? This question is particularly relevant for considering the prospect of increasingly adaptive agent-based technologies on a more general level.

3.6 DISTRIBUTED EPISTEMIC AGENCY

The idea of adaptive agents or multi-agent systems opens the door to visions and promises of new kinds of technologies that will fundamentally change the way humans relate to technology on a cognitive level. The association of the concepts of adaptive and self-organization with distinctly cognitive terms, such as learning, knowledge and decision making, inspires visions of agents that automatically organize and structure information and knowledge without humans having to fully specify how this should be done. At the far end of the spectrum are those visions that foresee technology advancing towards "self-organizing knowledge structures" within one global cognitive system, in which it does not matter much where one wants to locate epistemic agency (Gershenson & Heylighen, 2003). The philosopher and cognitive scientist Clark, for instance, reinvents Licklider's dream of human-

computer symbiosis, when he enthusiastically professes a new phase in the human/technology relationship instigated by the advances in 'user-sensitive' technologies:

> New waves of almost invisible, user-sensitive, semi-intelligent, knowledge-based electronics and software are perfectly posed to merge seamlessly with individual biological brains. In so doing they will ultimately blur the boundary between the user and her knowledge-rich, responsive, unconsciously operating electronic environment. More and more parts in our worlds will come to share the moral and psychological status of parts of our brains.
>
> *(Clark, 2003, p. 34)*

Clark foresees the advent of increasingly smart and adaptive technologies that will "learn" about humans and "dovetail" to their activities (p. 20). His vision is illustrative of the type of agent discourse that builds on an enticing but abstract, quasi-evolutionary narrative of a forthcoming shift in the human/technology relationship.

Clark's use of metaphors such as learning and adapting leads him to make claims about the changing ontological status of these technologies. He envisions future human/technology relationships to be so tight that humans will unconsciously come to expect and trust the input of agents, experiencing them as "natural extensions" of their personalities and cognitive abilities. For Clark, the similarities between humans and technologies will no longer be merely metaphorical. When future technologies "dovetail" back, he argues, technologies and humans will come to share a "moral and psychological status". Indeed, he professes that humans and technologies will function in such "intimate harmony" that it will serve "no legal, moral or social purpose" to draw a line between the two (p. 30).

In one big leap Clark goes from current development efforts in which 'adaptive system' serves as a design metaphor to equating future humans and technologies in terms of their competences and agency. Such a vision suggests that designing computer systems *as if* they were adaptive systems will enable these systems to interpret information, generate 'new' knowledge, and make decisions in a similar way as humans do. Existing constraints, boundaries and asymmetries will magically dissolve. Regardless of whether technologies can be made to simulate the adaptive or learning competences of humans, Clark's vision of future human-technology relationships exemplifies how extracting metaphorical concepts from the context in which they are meaningful can result in the

attribution of qualities to technologies that come in through the back door.

Clark's vision follows from his *Extended Mind* theory that builds on the turn in cognitive science towards a situated and embodied perspective of the mind and cognition. He considers humans to be *natural born cyborgs*: technology is already such a significant element in human practices that we humans would not be the beings we are without them. Like Hutchins, Clark looks upon the human/technology relationship as constituted by a continuous process of 'looping interactions' between humans and cultural and technological environments that extends over time and place, through which each element constitutes and shapes the other. Future 'dovetailing' technologies, according to Clark, will increase this intimate coupling.

Because of the prominent role of technology in cognitive processes, the question of what is a tool and what is mind becomes problematic for Clark. The hippocampus and the frontal lobe are as much a tool as a pen and paper. "It is tools all the way down" (p. 36). The idea of a conscious self in charge of operations, to him, is an illusion that distracts from the real subject of analysis. He suggests a reconceptualization of the mind (rather than cognition) as *extending* beyond the physical brain to include the external technological "props and aids" we recruit and exploit to structure and scaffold our reasoning (Clark & Chalmers, 1998). In contrast to Hutchins, for Clark this reconceptualization has further ontological implications that reach beyond providing an analytical framework for the study of cognitive processes in practice. He concludes that we cannot pinpoint the locus of the ultimate decision making and control of our behavior in the many neural biological and non-biological structures involved in reasoning. He thus (re)conceptualizes not only cognitive processes as distributed across humans and technologies, but also epistemic agency. In focusing on dissolving boundaries, however, Clark overlooks some significant differences between humans and technologies. A discussion between Hutchins and Bruno Latour highlights these differences.

In STS, actor-network theorists like Latour have also challenged the notion of epistemic agency as a property of humans that distinguishes them from non-humans (Latour, 2005). In his review of *Cognition in the Wild,* Latour applauds Hutchins' attempt at a symmetrical treatment of humans and the world (Hutchins et al., 1996). The theory, Latour contends, offers the tools to view cognition as a process in which no reference has to be made to human agency. "Thinking becomes an

ingenious way of constantly shifting from one medium to the other until one reaches 'simpler' or 'easier' tasks by delegating more and more tasks to other actors in the setting, either humans or non-humans" (p. 57). From this point of view, new intelligent technologies would do nothing more than internalize more cognitive processes and skills. No special properties currently only available to humans are transferred to technologies or are restricted to human beings. In fact, according to Latour humans do not have special species-defined properties. The disappearance of distinctions between humans and technologies, from Latour's point of view, is more a consequence of reconsidering conventional beliefs about the nature of humans and technologies and what separates them, than of adding abilities to technologies.

Nevertheless, some relevant differences between humans and technologies still remain that have to be accounted for.[17] As Latour to his dismay points out, the symmetrical treatment of humans and non-humans is not carried all the way through by Hutchins. In the Distributed Cognition point of view as presented by Hutchins, humans and their technological counterparts are differently situated in these systems. Hutchins assumes human agency - located within an individual or distributed across a group - to be the driving force behind the alignment and coordination of representational media to enable the propagation of representational states. "The thinker in this world is a very special medium that can provide coordination among many structured media – some internal, some external, some embodied in artifacts, some in ideas, and some in social relationships" (Hutchins, 1995, p. 136). He attributes to humans the special and exclusive status of ultimate coordinator of the components within cognitive systems. In reply to Latour, Hutchins remarks that his aim is to challenge old boundaries, but not to erase what lies inside them. "The work must be done somewhere, and some of the work will be done in regions that lie inside the bounds of persons" (Hutchins et al., 1996, p. 65). Hutchins' remark underlines that a reconceptualization of cognition as distributed across heterogeneous systems deconstructs the conceptual differences between humans and

[17] Latour recognizes the significance of differences, but differences should be treated as an effect rather than a given. In an article co-authored by Michel Callon, he writes: "Since differences are so visible, what needs to be understood is their construction, their transformations, their remarkable variety and mobility, in order to substitute the mobility of little local divides for one great divide. We do not deny differences; we refuse to consider them a priori and to hierarchize them once and for all" (1992, p. 356).

technologies as defined by the traditional notion of cognition, but it does not make the processes within the individual components the same.

Even in Clark's "complex reciprocal dance" of mutual-creation, technology and humans are not treated equally. In arguing for his theory of 'extended minds', Clark implicitly places epistemic agency at the human side of the cognitive system. The "plasticity" and "opportunism" of the human brain, according to Clark, is what enables the shaping of cognitive and cultural environments, as humans create and adjust to technological props and scaffolds (2003, Chapter 4). Clark's vision of 'new waves' of electronics and software underlines this asymmetry. His 'dovetailing' technologies present a phase shift in the intertwining of human beings and technology. If Clark argues that the mind already extends beyond the 'skin-bag', why does he feel the need to profess the ability of future technologies to increase this apparently already snug fit? For Clark, too, there still seem to be some relevant differences between humans and technologies. Yet, the significance of these differences disappears from view as a result of a preoccupation with deconstructing the boundaries erected by conventional cognitive theories.

Although, as Latour points out, Hutchins' human-centered stance is an inconsistency in carrying through a symmetrical treatment of humans and technologies in his analysis, it reflects what Suchman calls a "durable asymmetry among human and nonhuman actors" (1998, p. 11). She points out that "analyses that describe the active role of artifacts in the configuration of networks generally seem to imply other actors standing just offstage for whom the technologies act as delegates, translators, mediators; that is human engineers, designers, users, etc." (ibid.). She suggests that this persistent presence of human actors is indicative of culturally and historically constituted differences among humans and technologies. Hutchins' thinker as a special medium is an example of such a durable asymmetry. His human-centered stance illustrates the intuitive and firmly rooted idea, prevalent in the cognitive systems which he analyzes, that humans are the ultimate authority and reference point. Although technologies may be part of cognitive processes, humans are positioned differently in these processes. Such asymmetries are indicative of the multiple discourses in which the metaphors pertaining to cognitive processes acquire meaning and which they, in turn, help to shape.

The ambition to challenge conceptual boundaries erected by cognitivist traditions exemplifies the transfer of metaphorical concepts from one context to another. Like abstraction tools used by the designers of technological systems these metaphors provide analytical tools that

highlight and hide aspects of the world. They are taken from a different domain in which they are linked to a network of concepts. Ronald Giere notes that the idea of knowing is traditionally associated with the notion of a conscious epistemic subject, i.e. "the thing that knows" (Giere, 2002, p. 642). With regards to the notion of distributed cognition, he argues that reconceptualizing cognition as a distributed process does not require us to apply associated concepts, such as consciousness or agency, to the system encompassing both humans and technologies. He states:

> Cognitive systems are, of course, human creations, products of human agency. But we can refrain from ascribing agency to anything other than the human components of such systems. Nor need we endow such systems as a whole with knowledge, belief or any of the other mental states we associate with individual human minds, particularly not with consciousness. The reason for calling these systems cognitive systems rather than, say transport systems or agricultural systems, is that they produce a distinctly cognitive product, knowledge. But without the human interaction, there would be no knowledge, just a complex physical process.
>
> *(Giere 2002, p. 644)*

Humans can still be regarded as entities in cognitive systems that come to know the result of these processes, according to Giere. Knowledge produced by machines is only regarded as intelligible, if sense can be made of it by the human component in the system. An automated theorem-proving program is successful in proving theorems, if the proofs can be understood and accepted by humans. The proofs have to fit into existing human knowledge systems. Giere's observation signals the privileged position of humans in existing practices. In these practices, concepts like agency not only serve a descriptive purpose. As I will argue in the next chapter, they have a normative force that affects the configurations of humans and technologies. If, as suggested by Suchman, these asymmetries originate in a historical and sociocultural context, they can change. The question that arises is what are consequences of such a change?

3.7 CONCLUSION

An important reason for the continuing attractiveness of the symbiosis metaphor is the malleable and imaginative concepts it builds on. Concepts like adaptivity and self-organization support an understanding of the relationships between humans and technologies as a harmonious and seamless cooperation between two entities. Agent advocates building

on this metaphor have positioned agents as entities capable of independently adapting to complex dynamic environments, and of supporting humans without being told what to do. Although (re)conceptualizing humans and technologies as two entities interlinked in a symbiotic system provides a powerful framework for understanding certain phenomena, it has its limitations. Such a conceptual leveling places the emphasis on defining similarities, while asymmetries are pushed to the background. It constitutes an abstraction that captures only small part of the connections between humans and technologies. The danger of this conceptual leveling is that it can result in gratuitous comparisons between humans and technologies, which can in turn lead to awkward conceptual leaps.

The discussion in this chapter underlines that we should be careful when we take metaphorical concepts 'out of context'. Concepts like adaptive systems or self-organizing systems are meaningful within constrained discourses, constituted by humans, technologies, practices, ambitions, rhetoric, and conceptual frameworks. Moreover, they serve particular purposes in these discourses. As design metaphors and abstraction tools, these concepts are instrumental in positioning research projects, in guiding the search for new forms of computer technologies, and in conceptualizing complex problems and system design. In the Combined project, the metaphor of adaptive system supports a design vision to guide research and integrate the various sub-projects, while the image of adaptive actor/agents communities helps to distinguish this project as an innovative, integrated solution to the development of crisis response technologies. The DPN researchers enlisted this concept to describe and structure their thinking about a more generic approach to information fusion. Their particular and narrow understanding of adaptive systems is directly related to discussions about the limitations of centralized, hierarchical control in information fusions systems.

Although, on closer inspection the interpretations of metaphorical concepts like 'adaptive systems' can diverge significantly, these interpretations cannot be completely separated. The metaphorical concepts discussed in this chapter are part of multiple interrelated discourses. Using these concepts to describe (future) technologies without making reference to the conditions under which these concepts are meaningful, can therefore lead to misplaced expectations. Rhetoric on the promises of increasingly adaptive computer systems supports the idea that leveling the competences of humans and technologies is an unproblematic solution to the problems that humans are confronted with in an

increasingly complex and information-rich world. The suggestion that future agents will be capable of operating in complex dynamic environments can lead to the mistaken belief that these envisioned agents can serve as a substitute for humans. Explicating the conditions under which a system can be understood as adaptive, therefore, facilitates the communication between designers, users, managers and possibly even policy makers.

As we saw, decontextualized accounts of computer systems that gather, interpret, filter and act on information in complex and dynamic environments independent of human coordination obscure the role of humans that design and work with these systems. In developing their approach to information fusion systems, the DPN researchers leave unexplored how these technologies affect human performance and the organization of sociotechnical systems in which decision-support systems operate. Their abstract representation of the problem domain distracts attention away from the problems of automation that continue to challenge designers of information technology and automated systems. Increased automation and unpredictable behavior system behavior can, for example, impair the ability of human operators to maintain situation awareness. The theory of Distributed Cognition and literature from Human Factors research demonstrate that automation is almost never an issue of replacing or substituting human actors. It changes activities, as it redistributes skills and responsibilities between humans and technologies, and imposes new cognitive demands.

The narrowly defined concept of adaptive system can be a valuable heuristic or powerful metaphor to guide the design of and structure our ways of thinking about the development of computer systems. However, the pragmatic goal of developing technologies for use or application requires a critical reflection on how the interpretation of these metaphors relate to other relevant conceptual frameworks, in addition to a broader contextualized analysis of human/technology relationships.

The malleability and context-specific nature of metaphorical concepts present a number of questions about the conditions under which analogies between humans and technologies can be usefully elaborated. In case of decision support, for example, questions that need to be addressed are: In what sense can adaptive computer systems best support humans in their tasks? What parts of the system should be adaptive and in what way? What are the benefits and limitations of conceiving of a (computer) system as adaptive and for whom? How do various discourse-specific interpretations of particular concepts differ

and what is the significance of the discrepancies? If the goal is to make sociotechnical systems more adaptive, then a decision-support system that automatically adjusts to contingent circumstances might not be the preferred solution. More 'flexible' and complex systems can actually inhibit the adaptivity of sociotechnical systems, when decision-support systems behave in unexpected ways. From the perspective of DC, the adaptivity of sociotechnical systems cannot be reduced to features of individual components. Consistent behavior and predictability, from this point of view, can be a constructive feature of technological systems that support a complementary coupling between humans and technologies.

Finally, taking metaphors out of context can lead to conceptual leaps that overlook historically and culturally constituted asymmetries between humans and technologies. Conceiving of humans and technologies as interlinked in one cognitive system provides a powerful framework to conceptualize and study cognition, action and human/technology relationships. It enables the study of cognition as a phenomenon that is not restricted to processes inside the human mind. Hutchins' theory of DC illustrates that such a framework brings into view cognitive activities that have been largely overlooked by traditional cognitive sciences. The theory illustrates an alternative interpretation of the proposition that humans and technologies are conceived of as entities that can be described in terms of general principles or mechanisms, as it presents humans and technologies as components in cognitive systems, linked through computational processes. This can be a useful approach to explain certain phenomena or to challenge existing theories and conceptions. However, this framework too is constituted by metaphorical concepts that highlight and hide particular aspects of human/technology relationships. The affective and emotional dimensions of these connections, for example, are excluded from the theory of DC.

By redefining concepts such as adaptivity and learning as properties of an extended cognitive system they are detached from the associations they have in broader discourses. These reinterpreted concepts can therefore not be transferred unproblematically to make claims about the ontological status of future agent technologies. Some persisting asymmetries remain within the organization of sociotechnical systems, which are indicative of the formative role of concepts like 'epistemic agency'. A contextualized analysis of metaphorical concepts draws attention to the significance of these asymmetries. Why do particular asymmetries exist? What is their significance? How do they influence the organization of sociotechnical systems? And what can conflicting interpretations of

metaphorical concepts tell us about these contexts? The next chapter will take a closer look at persistent asymmetries and their significance with regard to the ambition to dissolve the boundaries between humans and technologies.

4. LIMITS TO THE AUTONOMY OF AGENTS

To what extent can or should artificial agents be autonomous? This is a central question in debates about the possibilities and risks of artificial agents. The ability to operate without continuous direction of human operators is an appealing feature of computer-regulated systems that perform tasks that are too complex, too dangerous, or that require accurate time-critical control. However, the prospect of integrating and incorporating increasingly autonomous technologies in daily practices raises concerns about the distribution of responsibility and accountability (Kuflik, 1999; Nissenbaum, 1994). Independently-acting, opaque, computational entities mask decision-making processes that are no longer directly traceable to or comprehensible for any single human. These concerns have become the focus of considerable attention in the artificial agent community. Agent advocates have suggested that the increasing complexity of computer technologies demands the design and implementation of *autonomous moral agents*. Such agents would have to be capable of reasoning about the moral and social significance of their actions (Allan et al., 2000). In the field of computer ethics, the suggested advent of 'truly' autonomous artificial agents has reignited debates about extending the class of autonomous moral agents to include artificial agents, in addition to humans (Allan et al., 2000; Floridi & Sanders, 2004; Stahl, 2004). The aim of this chapter is to reconsider the concerns raised by the idea of autonomous artificial agents from the perspective set out in the previous chapters.

In the previous chapters I have challenged the idea of leveling humans and technologies as an inevitable, desirable, or 'logical' outcome of technological development. I have argued that a relational and contextualized perspective on the connections between humans and technologies shows a much richer space of possible human/technology configurations. A preoccupation with a blurring of boundaries between humans and technologies obscures the view from the complementary roles of humans and technologies in these configurations. Moreover, I have argued that the proposed conceptualizations of artificial agents and the metaphorical concepts that they build on should be considered within the context in which they acquire meaning. Extrapolating these concepts from their context of use can lead to visions of future societies that make awkward conceptual leaps. In addition, it clouds the view of the

significance of historically and culturally constituted asymmetries between humans and technologies. The central focus of this chapter is on the conflicts that result from the confrontation of different meanings and roles of concepts. In particular, the discussion concentrates on the interplay between the moral aspects of human/technology relationships and the use of the metaphor of autonomous agents.

An account of autonomous computational agents should begin by asking how we can meaningfully speak about these technologies. Given the metaphorical, context-dependent and instrumental character of concepts used to support visions about future agents, we cannot take claims about their autonomy at face value. Despite several attempts in the agent literature to find an all-inclusive definition, autonomy remains an elusive and ambiguous concept, much like the term agent itself (Franklin & Graesser, 1997; Nickles et al., 2004; Wooldridge & Jennings, 1995). A closer look at the concept of autonomy in agent research, in the first part of this chapter, reveals that the tension between optimistic promises of autonomous artificial agents and the various concerns that they produce arise from the confrontation between two different conceptions of autonomy. On the one hand, autonomy is a concept inextricably linked with the notion of a moral and rational person rooted in a liberal democratic tradition. On the other hand, as inherited from the cybernetic roots of the computer science discipline, autonomy is a measurable and observable property of the relationship between biological or mechanical systems and their environments.

The tension between the two conceptions indicates that a persisting asymmetry between humans and technology remains. This asymmetry leaves the human as the ultimate morally responsible party and implies a preference for particular human/technology relationships. After exploring this asymmetry in more detail, the final part of this chapter looks at how the ambition to level humans and technologies, by developing moral agents or through a conceptual levelling, can affect these relationships. I argue that rather than looking for overarching solutions, the tension should be addressed at a local level from a sociotechnical perspective. Rather than asking whether technologies can or should be moral agents, we should be concerned with how different interpretations of autonomy can shape human/technology configurations in different contexts.

Since 2003, a piece of software onboard the Earth Observing 1 (EO-1) satellite is deciding which salient scientific events on earth - e.g. volcanic eruptions, flooding or ice break-up - the satellite should be paying attention to (Chien et al., 2005).[1] The software onboard the EO-1 is called the *Autonomous Science Agent,* and is developed as part of the *Autonomous Sciencecraft Experiment* (ASE). Given particular high-level goals, the onboard software collects, analyzes, and reacts to science data on its own. It uses machine learning and pattern recognition algorithms to scan images of the earth for interesting anomalies. It will, for instance, look for changes in volcanic activity by comparing observations. Based on the results of this analysis, the software can re-plan upcoming mission operations, and execute these re-planned responses. The ASE team built this system to enable "autonomous goal-directed exploration and data acquisition to maximize science return" for planetary science, space physics and earth science (NASA, 2006). Human operators in the control center only have a limited set of opportunities to instruct earth observation satellites to take pictures of the planet, because of the physical constraints on communication with these spacecrafts. This makes it difficult to study for instance 'short-lived science events' (such as volcanic eruptions, dust storms, etc.). The ASE team therefore set out to develop self-flying and self-governing spacecrafts that can operate for extended periods without human intervention and make decisions about observation goals.

The ASE follows in NASA's long and diverse track record in pursuing the development of autonomous robots that can operate at considerable distances from earth with minimal human direction.[2] Manually maneuvering and operating robots on other planets, such as the Spirit and Opportunity rovers on Mars,[3] is a very time-consuming process, as

[1] The ASE software continues to operate onboard the EO-1. NASA provides updates on the mission status of the ASE on (http://ai.jpl.nasa.gov/public/projects/ase/status.html).

[2] NASA's competitions exemplify its interests in autonomous technologies. As part of NASA's Centennial Challenge - a program of contests to stimulate innovation and competition in solar system exploration and ongoing NASA mission areas - the agency offered $250,000 to "develop technologies enabling robots to perform complex tasks with minimal human intervention", such as building structures, as well as to "design and build autonomously operating systems to excavate lunar regolith, or 'moon dirt', and deliver it to a collector" (http://exploration.nasa.gov/centennialchallenge/cc_index.html).

[3] See http://marsrovers.jpl.nasa.gov/home/

the communication between the human operator and the rovers suffers from serious delays. In certain time intervals communication is not even possible. Furthermore, controlling these robots is a challenge, because human operators can only rely on very small amount of sensory information, such as camera images under a limited viewing angle (Woods et al., 2004). This makes repair and recovery of remotely operated systems an impractical and costly enterprise. The idea of intelligent agents that automatically and independently operate and control robots and spacecrafts in dynamic environments suggests an appealing solution to overcome the limitations of manually operated spacecrafts.

The objectives behind NASA projects like the ASE are illustrative of the motivations that drive agent researchers to develop autonomous agents. The notion of autonomous computer systems has inspired optimistic visions of artificial agents that will be capable of replacing or supporting humans in an increasing number of tasks. The analogy with human autonomy supports a prevalent rhetoric in agent discourse of future worlds, in which computers will become animate entities that independently set out to accomplish their own goals, as if they have a life of their own. At the same time, this rhetoric presents future artificial agents as *delegates* or *collaborators* of humans. These agents will go out into complex information networks and physical environments to perform tasks 'on behalf' of humans. Personal digital assistants will, for instance, manage our daily appointments and our communication with others (Aarts, Marzano et al., 2003; Maes, 1994a). Social robots will take care of medical patients and our elderly (Dautenhahn, 2002; Fong et al., 2003). Military autonomous vehicles will take the place of human soldiers and go out into combat (Arkin & Moshinka, 2007). These visions build on the age old and persisting dream of building technologies that will relieve humans from their burdening tasks, and improve efficiency and safety.

The idea of autonomous artificial agents does not only lead to optimistic promises. The prospect of increasingly autonomous technologies arouses anxiety about technology 'out-of-control'. When the internal decision-making processes of technological systems are so complex that humans can no longer comprehend or intervene in these decisions, humans will be left at the mercy of these machines (Joy, 2000). It brings to mind 'doomsday scenarios' of the kind explored in movies like *Dr Stranglove* and the *Terminator*, in which the surrender of control to complex computer systems has catastrophic consequences for human life. Such dystopian images resonate in the responses to the recent surge

in social robotics research and the associated discussions on the ethical aspects of human-like independent robotic entities in various domains of society.[4] The advances in robotics have set in motion a range of initiatives in Europe, Japan and South Korea to formulate preemptive codes of ethics, protocols and even legislation that specifically address these aspects.[5] For the most part, these initiatives focus on the "human ethics" of designers, manufacturers and users (Veruggio, 2006, p. 27). In other words, they offer guidelines and principles for the design and use of robotic technologies, with respect to such values as safety, privacy and reliability. Nevertheless, in the popular media such initiatives invariably lead to the association with the *three robotic laws*, featured in Isaac Asimov's fictional stories.[6] These robotic laws exemplify the notion of endowing robots and computers with some form of moral or social knowledge, or 'robot ethics', to ensure that future autonomous technologies will adhere to human values and norms. Recent discussions in computer ethics on artificial moral agents show this association is not exclusive to popular discourses (Allan et al., 2000).

An abstract and decontextualized notion of autonomy runs the risk of confusing the different meanings of autonomy that it acquires in different contexts. This can obscure the role that the varying interpretations of autonomy play in the configuration of the connections between humans and technologies. The result is a restricted view of the questions and choices posed by the development of autonomous technologies. In discussing the social and moral aspects of autonomous agents, it is helpful to distinguish between two conceptions of autonomy. These

[4] The *Sunday Times*, for instance quoted the roboticist Professor Ronald Arkin as saying: "The question is what authority are we going to delegate to these machines? […] Are we, for example, going to give robots the ability to execute lethal force, or any force, like crowd control?" (Habershon & Woods, 2006).

[5] See for instance the BBC new coverage of South Korea' initiative ("Robotic age poses ethical dilemma", 7 March 2007, http://news.bbc.co.uk/2/hi/technology/6425927.stm) or the New Scientist Tech article ("South Korea creates ethical code for righteous robots", 8 March 2007, http://www.newscientisttech.com/article/dn11334-south-korea-creates-ethical-code-for-righteous-robots.html). Another similar initiative is the RoboEthics Roadmap released by the European Robotics Research Network (Euron) (http://www. roboethics.org/).

[6] Asimov introduced the following three laws in his short story *Runaround*: 1) a robot may not injure a human being or through inaction, allow a human to come to harm; 2) a robot must obey the orders given it by humans 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 Laws (Asimov, 1950).

conceptions are rooted in distinct conceptual systems that are confronted in the agent discourse.

Two conceptions of autonomy

Literally, autonomy means self-governance or self-rule, i.e. the ability to act independently of external direction. On the surface this seems relatively unproblematic. However, as a concept firmly rooted in contemporary Western society it has some noteworthy added connotations. Although originally, ancient Greek philosophers used the term - a composition of 'self' ('autos') and 'rule' or 'law' ('nomos') - in reference to states, throughout the ages it has become increasingly associated with the idea of personhood. As a defining feature of a person, it is a concept that serves to denote the capacity that most people like to think they possess, i.e. the ability to make our *own* decisions based on our *own* authentic independent motivations and desires (Christman, 2003a). We tend to ascribe to ourselves the ability to reason and make judgments independent of external forces or constraints (e.g. other people telling us what to do). We are insulted or frustrated when we are constrained in this ability.

It is largely due to Immanuel Kant that the concept of autonomy has come to refer to a property of persons. He saw the need to elaborate the concept of autonomy to substantiate his account of the validity of moral rules. From a Kantian perspective, a person is autonomous if he or she acts according to *universal moral rules* or *principles* (Hill, 1989). An autonomous person is not directed in his or her actions by external or internal influences, such as consequences or desires. He or she acts for moral reasons and because it is an objective obligation to do good. It is because an autonomous person is necessarily a *rational* being, that he or she can *rationally* determine which moral principles are authoritative. This conception of autonomy does not refer to the *freedom* an individual has to make choices, but to the ability a person has to rationally *will* to do what is objectively good. The implications of the autonomy of a person are that he or she deserves respect and is to be treated and judged as a moral agent. With Kant, therefore, the concept acquired an explicit moral connotation strongly tied to the notion of rationality.

Kant's moral theory and his notion of autonomy have had a significant bearing on moral, social and political thought that shaped contemporary Western society. In his analysis of the concept of autonomy in moral and political philosophy, John Christman notes that the idea of autonomy is an essential element of liberal democratic traditions, as they are founded upon the notion of a person as autonomous agent

(Christman, 2003b). In these traditions a 'normal' person is *assumed* to have the capacity for autonomy, and should not be significantly inhibited in her condition to exercise this capacity. A person has the right and the obligation to act as an autonomous agent. As such, autonomy is expressed in the foundations of our legal system, in human rights, but also in our daily treatment of other people. In our activities, we assume a person makes decisions based on independent thought processes, and excuse or fault them when they fail to do so.

Christman notes that "specifying more precisely the conditions of autonomy inevitably sparks controversy and invites skepticism about the claim that autonomy is an unqualified value for all individuals" (2003, p. 2). In practice attributing autonomy to humans on the basis of the fulfillment of a set of conditions turns out to be a less than straightforward endeavor. We attribute autonomy to persons in degrees. An adult is generally considered more autonomous than a child. As individuals in a society our autonomy is considered to vary because we are manipulated, controlled or influenced by forces outside of ourselves, such as by our parents or through peer pressure. Moreover, internal physical influences, such as addictions or mental problems, are perceived as further constraining the autonomy of a person. Critics have therefore challenged traditional conceptions of autonomy in liberalism. They claim that these conceptions fail to capture a person's embeddedness in a social and physical environment, or the complexity of identity and self (Christman & Anderson, 2005).[7]

Regardless of whether individuals really have the capacity or are actually given the right, in liberal-inspired Western societies, the concept acts as a fundamental organizing principle. It is a value that is deeply engrained in these societies. Autonomy associated with the idea of personhood is an ideal to strive for, and therefore not an innocent concept. As an ethical concept, ideal or organizing principle it has an impressive normative weight. It is instrumental in negotiating the boundaries between freedom and determined behavior, between moral and causal responsibility, as well as between humans and non-humans. As an assumed property of an individual 'rational' person, it provides the means or the conditions to presuppose that people reason and act

[7] In moral and political philosophy a wide range of different conceptions of autonomy has been offered. The philosopher Joel Feinberg, for instance, identifies four different meanings: the *capacity* to govern oneself, the actual condition of self-government and its associated virtues, an ideal of character derived from the virtues associated with the condition of autonomy, and the sovereign authority to govern oneself (Feinberg, 1989).

according to moral law, and to assign responsibility. It offers the conditions to blame or praise a person for her actions, because she is considered to be able and in the position to voluntarily decide upon acting.

A normative conception of autonomy can be distinguished from the concept of autonomy as an element in formal descriptions of *self-regulating* systems, as exemplified by the description of the Autonomous Science Agent. The characterization of the software onboard the EO-1 as an entity that operates with minimal human intervention is reminiscent of more conventional functional ways of thinking about *automation,* rooted in the cybernetic and systems theory traditions. Scholars working in these traditions continue to focus on the study of processes of organization and control in systems. They aspire to develop a general theory of how living organisms, machines and organizations can be explained in terms of the dynamic processes of communication and control within and between systems.[8] The level of autonomy, from this perspective, describes an abstract relationship of control between systems. In conventional ways of thinking about automation, it concerns the level of control that the machine has over the execution of a process, in relation to how much human intervention is required. Complete autonomy is on the far end of a continuous scale of increasing automation, where automation can be regarded as the use or introduction of machines or computers that are delegated tasks to complete without direct and continuous human control (Sheridan, 1992).

Automation usually involves a process of mechanization of tasks, where routine actions are translated into some formalized structure. It describes a particular relationship of control between humans and technologies, in which control over tasks is *distributed* or *delegated*

[8] Autonomy became an explicit part of the cybernetics vocabulary with the introduction of the idea of *autopoiesis*. Two cognitive biologists Humberto Maturana and Fransico Varela introduced the concept of *autopoiesis* in cybernetics to describe the self-organization of systems (Maturana et al., 1988). *Autopoietic* systems are self-producing units constituted by a recursive network of interactions between components that are produced by the network itself. An autopoietic system, like a biological cell, is closed and therefore autonomous because of its isolation in terms of organizational provocations. For computational systems this means that no predefined plan established by an external controller of how to act in a given situation is inscribed in the system. In cybernetics, a system is said to be *organizationally closed* if its internal processes produce its own organization. Interactions with the environment are possible in the form of input and output, but they do not control the internal organization; rather the system transforms and reproduces its own internal organization while maintaining its identity in response to the dynamics of the space or environment in which it exists.

according to the capacities of both. One often cited and illustrative description of the roles of humans and technologies in automation is provided by MIT professor Thomas Sheridan, who has introduced a gradual scale of automation to illustrate the incremental *levels of control* that can be shared between human operator and computers (Sheridan, 1992). The minimal level of automation leaves it up to the human to make all the decisions and take all the actions; the computer offers no assistance. The higher the level of automation, the more the decision-making opportunities for humans are constrained by the actions of the computer, going from offering a set of complete decision/action alternatives, to offering a narrow selection of choices. The more a system is capable of collecting, analyzing, interpreting and acting on information - be it sensory information or explicit symbolic representations of knowledge - the more autonomous the system is considered to be. Higher levels of autonomy are, then, attributed to those automated systems (machines or computers) that are left to perform tasks on their own, and have the *authority* over these processes, i.e. humans have neither the need nor the ability to intervene.[9] The EO-1 satellite controlled by the Autonomous Science Agent is thus more autonomous as compared to a remotely manually controlled satellite. In this account of automation, the concept of autonomy is cleansed from its normative connotations. It describes an observable and measurable property of a relationship between entities. The notion that an automated technology can act for extended periods of time on its own, has no moral implications for the technology itself; it does not attribute to it certain rights or obligations.

Although the two conceptions of autonomy share a family resemblance as a relational property of an entity, it is against the background of these two distinct contexts that significant differences in meaning become apparent. Autonomy as an assumed property of a person supports a conception of the person as a rational, moral individual that still underlies our contemporary beliefs about what it means to be human. This concept is more than an element of an analytical description of human beings, as separate from other entities. It is inextricably tied to the social, economic, political, juridical, and ethical contexts in

[9] Note that conventional ways of thinking about automation are primarily concerned with the operation of a single machine in a closed environment. Cybernetics and systems theory have also influenced research on multi agents systems (see the previous chapter). However, MAS approaches extend the focus to systems encompassing multiple entities in open environments.

which a person acts. In these contexts, it becomes a non-quantifiable assumed property that serves as an organizing principle and has a prescriptive quality. In contrast, autonomy in cybernetic-inspired ways of thinking about automation is primarily a functional and descriptive concept that supports formalized models of a gradual scale of organizational dependence between two entities. It is an abstraction that serves an instrumental purpose in formulating mathematical and logical models to describe biological and artificial systems in general terms. The most notable difference between these two meanings of autonomy is that the first has strong normative connotations, whereas in the second the ethical dimension of the relationship of control between human and machines plays an inconsequential role. It is this difference that generates concerns when the decontextualized visions of increasingly autonomous agents are discussed. As the discussion in the next sections will show, the tension between the two conceptions of autonomy plays an important role in configuring humans and technologies. I will start by taking a closer look at the varying interpretations of autonomy in agent literature.

4.2 SELF-REGULATING AGENTS

In agent research, the concept of autonomy has been rediscovered, dusted off, and put to new use to denote one of the most characteristic features of agents (Nickles et al., 2004). It is the one feature of an agent, all descriptions seem to agree upon, that captures what an agent is in its most abstract form. It is certainly a pervasive term in this field of research, as is made evident by the generous use of the adjective in combination with the term agent in conference and journal names.[10] As a key feature it is also one of the most contested concepts. Nevertheless, a functional, descriptive conception of autonomy is predominant in agent research. Thinking in terms of autonomous agents allows agent researchers and developers to describe, in a high-level abstract sense, a computational entity (e.g. a software component or robot) as a self-contained, self-regulating, interactive unit that operates in some physical or digital environment. Humans and human activity play a minor role in a significant part of this research.

Indeed, a number of research projects on autonomous agents are driven by the explicit objective of taking the human 'out of the loop'. Writing about their research on NASA's Autonomous Scientific Agent,

[10] Examples include the yearly *International Conference on Autonomous Agents and Multi Agents Systems* and the *journal Autonomous Agents and Multi-Agent Systems*.

Chien et al. suggest that their software should enable the EO-1 to perform autonomously in at least two ways: it can fly and control itself in an unknown environment and it can perform a significant part of the expert scientific data gathering, analysis and interpretation tasks (Chien et al., 2005). The emphasis in the account of Chien et al. is not on analyzing the dynamics of the human/agent relationship; rather it is on isolating the actions that should be performed by the EO-1 from this relationship. The objective of the ASE team is to eliminate the requirement for a human operator in the sensing, processing, acting and control loops. The spacecraft (or rather the software onboard the spacecraft) can then locally define observation goals based on an automated analysis of satellite images and derive mission operation plans from these goals. These mission plans direct the sequence of actions of the spacecraft. The kind of autonomy pursued by the team takes the form of relative *closure* of the system's organization, in the sense that physical actions, decision making and information processing take place in a *closed control loop*, circumscribing only the spacecraft and its immediate environment. This conceptualization of autonomous agents is similar to the machines described at the high end of Sheridan's scale of automation.

More elaborate definitions of autonomy have been proposed in the agent literature to characterize agent-based systems as a different kind of computer technology. For most agent researchers, the concept of autonomy is one of a range of abstraction tools to delineate agents as a software engineering approach. The emphasis on the autonomy of agents in multi-agent systems, for instance, positions agent-oriented approaches within a spectrum of software engineering methods, including *Object Oriented Programming, WebServices* or *Distributed AI*. The use of the notion of autonomous agents reflects a particular way of thinking about how one part of a program relates to other parts (Luck et al., 2005).

Agent researchers Michael Wooldridge and Nick Jennings, for example, contrast agents as components in a software program to 'objects' (Jennings & Wooldridge, 1998; Wooldridge, 2002). Both agent-oriented and object-oriented approaches represent software programs as composed of discrete units that encapsulate methods and data. An object-oriented software program operates through message passing, where objects invoke methods upon one another. According to Jennings and Wooldridge, objects are thought of as having some control over their internal state, in the sense that this state can only be accessed or modified through the methods that the object provides. An external

entity cannot change the structure of the processes within the object. An object *x* can be 'told' what to do by another object *y* when *y* invokes a method *m* provided by *x*. The object *x* has no control over whether the method is executed or not. In multi-agent systems it cannot be taken for granted that an agent will execute a method when it receives a message to do so. Jennings and Wooldridge asserts that agents are conceived of as having control over their own actions, *as well as* over their internal state. They are not externally directed in the generation and completion of their goals or their decision-making processes by other agents. Unlike objects, agents are thought of as 'requesting' other agents to perform an action. The decision to act upon this request is left to the recipient. According to Wooldridge this distinction is summarized in the slogan: "Objects do it for free; agents do it for money*"* (1999, p. 35). Autonomy as a distinguishing feature of agents encourages researchers to explore alternative metaphors to characterize interactions between the units in computer systems, like 'cooperation', 'collaboration' and 'negotiation'.

Scholars concerned with the theoretical underpinnings of agent research have proposed a variety of definitions of autonomy. They have drawn on philosophical and sociological theories to explicate and formalize the mechanisms that enable artificial agents to generate and pursue goals independent of external factors (Verhagen, 2003). Luck et al. speak of explicitly *operationalizing* the term autonomy (Luck et. al., 2003). This entails formulating definitions of autonomy that specify in detail, and preferably in logical and mathematical models, perceived features of exemplary autonomous entities. The varying definitions highlight different aspects of autonomy, including independence from other agents in decision-making processes, the ability of an agent to generate goals from its own motivations, as well as the degree to which an agent can give itself laws (Elio & Petrinjak, 2005; Luck et al., 2003; Maes, 1994b).

Cognitive scientists and agent researchers Christiano Castelfranchi and Rino Falcone, for example, explore various kinds of autonomy that set agents apart from mere automation (Castelfranchi & Falcone, 2003, 2004; Falcone & Castelfranchi, 2001).[11] As part of their work on a sociocognitive theory of delegation, dependence and control, they

[11] An agent, according to Castelfranchi and Falcone, is more than an automatic entity, if it is a self-adapting system, "able to find its own solutions not only thanks to intelligence but also thanks to autonomous learning or evolution" (2003, p . 13). This form of autonomy allows an agent to perform independently of environmental influences.

distinguish 'social autonomy' from other kinds of autonomy like 'goal autonomy', 'executive autonomy' and 'autonomy from the environment'. They note that to understand the interactions between agents the concept of autonomy needs to be extended. It has to capture the relationships between cognitive (or intentional) agents in a social organization. In a social organization, external powers, such as conventions and norms, can interfere with the ability of an entity to act autonomously. In Castelfranchi and Falcone's view an agent is fully socially autonomous when it has its own goals, it can decide about these goals, it is not coerced into accepting goals from others as its own, and its goals and beliefs cannot be automatically modified or changed by outside factors.

In the context of discussions on system architectures, software engineering methods and the theoretical underpinnings of agent research, 'autonomy' features as a formalized conception of self-regulation, with a particular emphasis on the ability of agents to generate and pursue their own goals. This conception is instrumental in understanding and developing computer systems in isolation from their connections to humans, as it supports conceptual frameworks to study particular technical and formal aspects of computer systems.

Humans feature as prototypical autonomous agents in descriptions of operationalized notions of autonomy, but their complexities are filtered out in these abstractions. Autonomy as a heuristic in the development of agents-based systems serves to explicate and formalize what constitutes the control of technological systems over their internal organization. In other words, what should the computer system be capable of if there are no humans around? In this technology-centered conceptualization, agents and humans are regarded as two separate, but sporadically interacting systems. Human beings are reduced to peripheral elements of no significant relevance for the internal processes of the agent. The connotations of human autonomy that make it such a contested subject in philosophy, such as moral responsibility and personhood, are erased or play a minor role.

Although the idea of an autonomous agent is instrumental in guiding the search for technologies that can automatically regulate themselves within an environment, in practice technologies do not operate in isolation. They become part of human social organizations and culture, in which a notion of autonomy with moral connotations is prevalent. A functional notion of autonomy leaves unexplored how humans work with automated technologies (see Chapter 2 and 3). Even the EO-1

satellite remains connected in multiple ways to earth scientists, human operators, engineers and other human actors. After having described their "absolute theoretical viewpoint of autonomy", Luck et al. caution that "there is value in studying the general concept of autonomy, regardless of practical concerns, but we must also address ourselves to the practical issues" (2003, p. 20). These practical issues come into view when technologies are discussed in terms of their application. The tension between different conceptions of autonomy comes into play in these accounts.

4.3 THE HUMAN IN THE LOOP

The *delegation* of control to artificial agents puts them in a continuous dependency relationship with humans, as decisions have to be made that bear upon human actions, welfare, rights, and obligations. This is the point where that nagging feeling of loss of control starts to become an issue. Increasingly complex, seemingly independently-acting technologies trigger concerns about responsibility, accountability and trust. The idea of artificial agents that pursue their own goals raises questions about the extent to which a human user can or will trust these agents to perform tasks appropriately and 'on their behalf'. Scholars critical of techno-enthusiastic discourses about the promises of computer technology argue that the idea of autonomous agents encourages the user to attribute a kind of decision-making capacity to the computer that sits uncomfortably with the practical implementation of responsibility and accountability in daily life (Johnson, 2006; Nissenbaum, 1994). Progressively autonomous technologies would hide more and more decision-making processes from their human operators. What happens when things go wrong? As we saw in the previous chapter, the combination of humans and complex automated systems can create (unforeseen) vulnerabilities and risks, which presents questions about who or what is responsible for resolving or preventing these uncertainties.

According to the organizational theorist Charles Perrow, the vulnerability of a system will increase in conjunction with the level of complexity. With every addition to the complexity of a system the possibility for accidents grows (Perrow, 1999). In his analysis of the vulnerability of automation in sociotechnical systems, Perrow distinguishes between *linear* and *complex interactions* that can occur within a system. Linear interactions are predictable and visible and can be traced along the linear sequence of events that one action sets in motion. Complex interactions

are those interactions that cannot be realistically anticipated and lead to unfamiliar and unplanned sequences. They result from multiple dependencies or unexpected interactions with the environment, and are therefore not intended in the design (Perrow, 1999, p. 78).[12] As a second variable to characterize systems, Perrow introduces the level of *coupling,* where he distinguishes between *loose* and *tight coupling.* In loosely coupled systems there is *slack* or there are *buffers* between the components that make up the system. This means that the effects of the operations of one component - which can be material and immaterial as well as human and non-human - do not directly affect other components. Slack between the components has the advantage that when one component fails, other components do not directly suffer the consequences. In tightly coupled systems, as exemplified by automated computer systems, components are closely adjusted to each other with limited variance in space, time and logic. The advantage of these systems is that they are more efficient in terms of time, production costs, and accuracy. However, they are also more vulnerable to accidents, errors or other unintended harmful consequences for humans. According to Perrow, linear loosely coupled systems are easier to control, and less vulnerable to unexpected interactions. They are controlled by humans either directly, where the human operator is responsible for aligning the various components facilitated by the predictability of the linear interactions, or indirectly, where the human developer deconstructs or organizes the problem in a linear sequence of interactions to enhance predictability and reduce complexity.

For Perrow, the problem with conventional automation is that when systems are more tightly coupled and the human is taken out of the loop the buffers and safeguards are reduced. There is limited room for interventions and they are more susceptible to the occurrence of unexpected interactions. Loosely coupled systems require the flexibility and the intelligence that currently only humans can provide. In the space between components, human operators can make "fortuitous" interventions and adjustments without significantly disrupting the sequence of events. Perrow assigns to humans the exclusive ability to take *appropriate* action or make "fortuitous" interventions in the face of unreliable environments and unexpected events when given the opportunity and space to perform (i.e. when they are not governed by rigid rules and

[12] The qualifications of linear and complex should not be confused with the level of sophistication; they only describe how interactions take place.

regulations). Humans should therefore be placed firmly in the loop as the ultimate responsible party.[13]

Perrow's analysis highlights the paradoxical idea of delegating control to and trusting independent, flexible, and thus potentially unpredictable technologies, which are proposed as a solution for the short-comings of humans. It underscores the tension between autonomy as an abstract functional concept and a more common conception of autonomy in Western contemporary society. From the point of view presented by Perrow, creating complex self-regulating technologies and thereby relinquishing control over certain processes reduces the opportunity for human intervention and inhibits the human operator's understanding of the system.

The problematic aspects of tightly coupled complex systems present a challenge for researchers concerned with the development of autonomous agents. The increasing focus on the idea of *adjustable autonomy* in the field of agent research illustrates that a number of agent researchers recognize the problems of delegating control to autonomous artificial agents and focus on exploring this dependency relationship between humans and agents (Bradshaw et al., 2004; Dorais et al., 1998). Adjustable autonomy offers an alternative to direct and continuous operation of technology on the one hand, and black boxed automation on the other. This notion is part of the growing trend to model the human/technology relationship after the idea of humans and agents working in a 'team' (Sierhuis et al., 2003; Sycara & Sukthankar, 2006). In conventional computing a technology is rigidly assigned a number of tasks. In contrast, the metaphor of working in a team implies that task allocation is more dynamic and agents as 'team members' participate in the 'negotiations' about these allocations.

Most research projects concerned with 'adjustable autonomy' work with a functional notion of autonomy and assign ultimate control to humans. The driving motivation behind these projects is to develop technology that can operate independently and make decisions for extended periods of time, but leaving open the possibility of transferring control over decision-making to humans. Agents in such a system perform a limited set of decision-making tasks (e.g. rescheduling

[13] Human-centered approaches in HCI are generally based on the premise that humans should ultimately be in command of the behavior of the system, and the design of the system should accommodate this role of human operators. For another plea for placing humans in the loop see Charles Billings' paper on Human-Centered Intelligent Systems (1997).

meetings or ordering meals), on their own. An agent determines when it should act more autonomously and when it should place itself under external control, on the basis of varying factors. The agent can, for instance, consult the user when it is confronted with conflicting information, or when the expected level of utility of the agent performing a certain task is below a certain threshold. A critical research challenge in this type of project is modeling the conditions that enable a successful coordination between humans and artificial agents (Scerri et al., 2002).

Most projects in research on adjustable autonomy assume that the human operator should be able to reclaim control over decision-making processes. For example, in describing their effort to implement adjustable autonomy in a prototype agent-based system for disaster response, Shurr et al. note that:

> Allowing humans to make critical decisions within a team of intelligent agents or robots is prerequisite for allowing such teams to be used in domains where they can cause physical, financial or psychological harm. These critical decisions include not only the decisions that, for moral or political reasons, humans must be allowed to make, but also coordination decisions that humans are better at making due to access to important global knowledge, general information or support tools.
>
> *(2005, p. 198)*

Although there is room for agents to play an 'active' role in deciding on task allocation and delegation of control, in this approach the human team members remain the ultimate responsible party. The autonomy of computer systems is treated as distinct from the autonomy of humans.

Nevertheless, some agent researchers seem to diagnose the problem underlying concerns about increasingly autonomous technologies in terms of a gap between humans and technologies that needs to be bridged: current autonomous technologies are not sufficiently like humans. These researchers have drawn an explicit analogy with human autonomy as a moral concept in their visions of how artificial agents can solve the problem that increasingly complex technologies pose. In their discussions of the various approaches to build *artificial moral agents*, Allan et al. write "as artificial intelligence moves ever closer to the goal of producing fully autonomous agents, the questions of how to design and implement artificial moral agents becomes increasingly pressing" (2000, p. 251). The observation of Allan et al. echoes a more pervasive solution suggested in the agent community to resolve concerns about loss of control. In particular, one suggested strategy is to move agents closer to

humans by formalizing and digitizing those abilities or qualities that enable humans to be part of a social organization. This strategy entails endowing technologies with the abilities that allow them to collaborate on the basis of an understanding of shared norms and social rules.

The aforementioned Cristiano Castelfranchi is one advocate of the development of agents that are able to reason about the social and moral dimensions of their tasks. If artificial agents are to be embedded in a complex socio-cultural environment, he believes, they should be capable of understanding the mechanisms of what he calls "social order", such that they can effectively support human activity (Castelfranchi, 2003). They have to be able to 'understand' the informal processes in spontaneous and 'bottom-up' interpersonal relationships that give rise to social order. This, he claims, requires a 'formalization' of these informal dynamic processes, in addition to the formal mechanisms such as rules, regulations, protocols and legislation. Artificial agents should be capable of reasoning about morality, culture, and law:

> Especially within the intelligent and autonomous agent paradigm, I believe that it is both possible and necessary to model these typically human and social notions. In order to effectively support human cooperation – which is strongly based on social, moral, and legal notions - computers must be able to model and 'understand' at least partly what happens among the users. They should be able to manage - thus partially 'understand' - for example permissions, obligations, roles commitments and trust.
>
> *(Castelfranchi, 2003, p. 51)*

From Castelfranchi's point of view, the possibility for creating artificial agents on a par with humans is unlimited: if the process can be rationalized and formalized then agents can be endowed with the capabilities that enable them to take part in social organizations and take over increasingly more decision-making tasks. Such agents would be capable of replacing humans in taking up the slack in sociotechnical systems.

A problem with efforts to formalize the mechanisms of social order or other types of mechanisms in order to create autonomous moral agents is that such approaches tend to exclude existing orderings and fundamental beliefs about humans and technologies from analysis. In other words, they overlook the normative role of 'autonomy' in a broader discourse as a concept strongly tied to the beliefs about what it means to be human. The functional and descriptive conceptions of autonomy that these efforts build on place humans and artificial agents on the same level, assuming that if the mechanisms supporting social or

moral reasoning can be formalized then humans and agents are treated as equal agents.

The contentious idea of ascribing moral responsibility to computers illustrates the tension between the different interpretations of autonomy that efforts to build autonomous moral agents tend to overlook. In philosophy and ethics, an autonomous person (i.e. a person that acts voluntarily and on the basis of free will) has traditionally been one of the main preconditions for the ascriptions of moral responsibility, in addition to other conditions like causal responsibility, the freedom to act and the power to control. Echoing familiar critiques of the early project of AI, scholars in the field of computer ethics have argued that artificial agents cannot be ascribed moral responsibilities, because they can never have the capabilities that make humans moral agents, such as mental states, intentionality or emotion (Johnson, 2006; Kuflik, 1999). Although discussions on the validity of such ontological objections have brought to light various aspects of 'autonomous moral agents', they leave the interplay between different interpretations unaddressed. To explore this issue further, the next section concentrates on the normative role of moral responsibility as a decidedly human concept in shaping human/technology configurations. The discussion will show that there remain some persisting asymmetries between humans and technologies.

4.4 PERSISTING ASYMMETRIES

The acquisition and use of knowledge about moral behavior is problematic when it comes to machines. This is not only because fifty years of research into the possibilities of artificial intelligence have demonstrated the extent and difficulty of building such technologies; it is also because deeply rooted normative beliefs about the boundaries between humans and technologies in Western cultures favor particular human-technology configurations. In liberal democratic societies humans and technologies tend to be differentially positioned in the process of responsibility ascription, as a result of particular orderings in these societies.

One way to look at the ascription of moral responsibility is as a social process that serves the objective of blaming, praising, sanctioning, or rewarding someone to obtain a result (Stahl, 2004). Human Factors researcher Victor Riley asks, "with all the complexities surrounding human interaction with automation, and recognizing that automation can perform many tasks more precisely and reliably than human operators can, one may wonder why we don't just automate the operator out of the

process altogether?" (Riley, 1996). The answer to this question, Riley suggests, is that as long as we feel the need to blame someone when things go wrong we will assign a responsible human operator. His comment indicates a persisting asymmetry in the treatment of humans and technologies when it comes to ascribe moral responsibility. We can 'blame' a computer or hold it 'accountable', by replacing or modifying it, but the search for moral responsibility does not stop there. Chains of responsibilities are traced back to human operators, developers, managers, or even politicians: there is a bug that needs to be fixed, developers or users did not have enough training, or the impact of technology was not accurately anticipated. After all, what happens when the Scientific Agent fails to produce the correct images, or an autonomous e-mail filter deletes important messages from your boss? The tendency will be to hold those humans developing, using, integrating, and managing technologies ultimately responsible and accountable for these failures. In the end, a computer program can be changed, but cannot be sued.[14]

In their analysis of the delegation of control and action, the sociologist Harry Collins and the philosopher Martin Kusch highlight the asymmetrical treatment of humans and technologies (Collins & Kusch, 1998). They observe that humans delegate only a particular kind of actions to technology. Machines are only delegated actions that do not require an understanding of the social and cultural context to carry out the behaviors. Collins and Kusch conceive of actions and their meaning as intimately tied to a culture, or a *form of life*, as they call it in reference to Wittgenstein. A form of life, they state, is constituted by shared *formative* actions that distinguish a society from other societies. These actions are tightly intertwined with a common net of concepts shared by the members of the form of life, as "intentions are conceptual and because concepts provide guidance for actions" (p. 11). In contrast to a piece of behavior, an action is more than a reflex. An action, like parking your car, is a meaningful composite of a set of behaviors and sub-actions.[15]

---

[14] Responsibility is an important concept in political, legal and ethical discourse. In the literature on responsibility, various kinds of responsibility are often distinguished such as legal and moral responsibility (Coleman, 2005; Johnson, 2001). As I am primarily concerned with the practical role of the concept of responsibility in the development of technology, I consider moral responsibility as offering a basis for other kinds of responsibility.

[15] Collins and Kusch let the distinction between action and behavior coincide with the distinction between natural and social kinds. Both natural and social kinds have a self-referential component. But whereas in natural kinds the reference extends outwards to

Collins and Kusch distinguish between two kinds of actions: *mimeomorphic* and *polimorphic* actions. Mimeomorphic actions can be reproduced by an individual outside of a cultural context. This individual can mimic the behaviors that constitute the action without understanding the significance of these behaviors. The reproduction of the behaviors appears to reproduce the action to a member of a form of life, who understands the meaning of the action. This kind of action can be performed by humans as well as machines, they argue. A soccer robot does not have to understand the rules of the RoboCup[16] soccer game in order to play it, just like a traffic light system does not have to understand the significance of its behaviors.

Polimorphic actions, Collins and Kusch maintain, can only be performed by a member of a society or community (*polity*). They are "rule bound" actions, in the sense that there is a collectively constituted right and wrong way to perform the action. Yet, these actions cannot "be specified by listing the behaviors in terms of which it could be carried out" (p. 23). If the behaviors were to be copied by an outsider it would not be the "same" action, as a polimorphic action can be performed in many different ways. For example, the mimeomorphic action of riding a bike or driving a car can be described by a set of behaviors. A robot could be made to perform these behaviors.[17] In contrast, the polimorphic action of riding a bike through city traffic requires an understanding of its meaning in order to carry it out correctly. A set of rules cannot exhaustively describe the possible events that would enable appropriate action under the contingent circumstances in traffic. You need to understand how to appropriately deviate from the formal rules. In the same way, the spelling checker on my computer stubbornly continues to underline the term "polimorphic" with a red line as I am writing this chapter. It does not understand, as I do, that in the context of Collins' and Kusch's theory, it is actually a meaningful pun rather than a spelling

---

something that exists independently of the reference, a social kind exist solely by virtue of its reference to itself. The self referential component in a natural kind such as "mountain" consists of the collective agreed criteria that classify something as a mountain. A social kind, such as "money", is "exhausted by the self-reference" (1998, p. 23). "Actions are social kinds because the communities are able to recognize their behavioral instantiations" (p. 23). Waving is a greeting because it is collectively taken as such.

[16] RoboCup is an international initiative that organizes a conferences and competitions, where robotics groups can showcase the abilities of their robots on a number of challenges, the main event being a soccer competition (http://www.robocup.org).

[17] Murata Manufactoring in Japan created "Murata Boy", a self-balancing bicycling robot (www.marutaboy.com).

mistake. It takes someone who understands the social context and the significance of the action, to perform a polimorphic action correctly.

A crucial difference between mimeomorphic actions and polimorphic actions is the extent to which members of a form of life are *indifferent* to the variations in the way behaviors are carried out. An action becomes polimorphic at the point where the members of a form of life start to care about the variations in behavior. At this point the meaning of the behaviors becomes negotiable, and it takes a member of a form of life to participate in this negotiation. Polimorphic actions can be performed in many different ways, but their correctness or appropriateness, and in this sense their similarity is bound by the shared conceptual system of a form of life. Machines, according to Collins and Kusch, 'behave' and can therefore only simulate mimeomorphic actions, but not polimorphic actions, as they cannot understand the significance of variations in behavior. Mimeomorphic actions unlike polimorphic actions can be delegated to a machine, because we do not have to negotiate or build a shared understanding with the machine.

Although Collins and Kusch are not specifically concerned with morality or moral responsibility, their distinction between polimorphic and mimeomorphic actions illustrates the prevalent anthropocentric bias in Western contemporary societies. This bias is based on a Kantian-inspired conception of the human as an autonomous person. Collins and Kusch note that the relationship between the delegation of action and moral responsibility is complex and often orthogonal (p. 63). In other words, the delegation of action does not mean that moral responsibility is also delegated. From their point of view, a person can be held morally responsible because she is part of a form of life and capable of polymorphic action. Humans can act in undetermined ways that nevertheless fall within the limits of acceptable behavior. This conception of moral responsibility attributes an ability to humans that distinguishes them from mechanical, determined systems: humans can deviate from rigid rules and protocols to perform their actions in an *appropriate* way. In order to act appropriately in unpredictable situations, it is not only necessary to have a static model of what actions are appropriate; it also requires the ability to make judgments in contingent situations based on an understanding of cultural knowledge and the social context. This understanding, according to Collins and Kusch, can only be gained through "socialization".[18]

[18] Arthur Kuflik offers a perspective on how socialization is part of moral responsibility. He identifies a particular kind of moral responsibility, which he calls "moral

Collins' and Kusch's analysis of action could lead to a discussion about which human qualities enable socialization, such as emotion, mental states or intentionality, but it also indicates another equally problematic issue for agent researchers who aim to level humans and artificial agents. Besides having the ability to be part of a form of life, it is necessary to be *considered* an entity capable of making such judgments. A cycling robot that manages to navigate successfully through city traffic can be perceived as understanding the rules of behavior. Yet, if an accident would occur, this understanding seems to dissolve; it is the designers or other human actors that allowed the robot to operate in city traffic who will be held morally responsible. Machines are not accepted as entities that can share our concepts, beliefs, and goals. As long as a computer is not considered to be an autonomous entity in this sense and humans are, moral responsibility remains a human affair.

Regardless of whether or not humans or machines can be capable of something as mysterious as intentional behavior, what Collins and Kusch's dichotomy signals is a fundamental normative belief about what it means to be human, as well as about the role of technologies. Being part of the form of life that Collins and Kusch ground their dichotomy in requires that one is accepted as an autonomous person rather than a cog in the wheel. The exclusion of machines from this form of life is the result of the deeply rooted modern notion of a person that draws a boundary between humans as ultimate moral authority and machines as not capable of understanding cultural norms. Enhancing the competences of technologies does not by itself dissolve this anthropocentric bias, as it is precisely this bias that sets constraints on the behavior of technological artifacts. Moral responsibility favors particular human/technology configurations in which technologies are conceived of and positioned as instruments that do things for or on behalf of humans. They are a means to an end. Such configurations are accompanied by requirements for controllability and predictability, which constrain (but not necessarily determine) the range of behaviors technologies can carry out.

Collin's and Kusch's perspective on the delegation of action highlights that the process of automation does not simply produce a

---

accountability responsibility". He argues that we consider individuals morally responsible agents if they cannot only give an explanatory account of themselves, but can also engage in a discussion about the appropriateness of their comportment and are willing to acknowledge, apologize and make amends for their possible errors in judgment (Kuflik, 1999).

substitute for human activity. It seldom involves a one-to-one mapping of tasks previously performed by human beings onto formalized mechanical structures. To ensure that the actions of automated technologies adhere to the goals and expectations of humans involved with the technologies, tasks, and responsibilities are deconstructed, reassembled, and reassigned. This process is visible in the development of the previously mentioned Autonomous Science Agent. The closure of the control loop in this automation, from Collins' and Kusch's perspective, comes down to redefining the actions to be delegated to the technological artifact in terms of mimeomorphic actions. The ability of the ASE to regulate itself is the result of a careful crafting of the various components that make up the system. The team working on the ASE spent a good deal of time verifying, testing and fine-tuning the models of the spacecraft to ensure that it would operate within the limits of acceptable behavior, as high stakes were involved in the Earth Observing 1 project. The failure of an EO-1 mission would have been very costly in terms of time and money.

Cichy et al. detail the meticulous process of validating the behavior of the EO-1 spacecraft equipped with the ASE to ensure that the software correctly encoded the operations and safety constraints of the EO-1 (2004). They describe a period of years of knowledge engineering, simulation, extensive testing and model review. "Any inaccuracies in these models could lead to ASE failing to achieve science objectives, or in the extreme, issuing unsafe sequences of commands" (Cichy et al., 2004, p. 3). Another paper recounts how the ASE team developed its models through an iterative process starting from a high-level action, such as "science observation" and "spacecraft pointing", working towards a full specification of the allowed behaviors of the system that were "consistent with existing ground operations and constraints of the EO-1 spacecraft" (Chien et al., 2005, p. 41). To generate a model, the team studied the sequences of commands involved in 'multi-activity objectives', such as calibrating instruments and collecting scientific data. Furthermore, a team of engineers, considerably experienced with working on the spacecraft, was employed to reason through all possible errors or contingent circumstances that could occur and verify whether there were any "incorrect parameters" or assumptions represented in the models. The resulting list of potential hazards then served as a basis for the design of appropriate safeguards. As ultimate safety measure the EO-1 operations team can disable the ASE commanding path or the ASE control of the EO-1.

The EO-1 example illustrates that within a context where humans are considered to be the ultimate moral (and epistemic) authority, the successful delegation of tasks takes place in physical, conceptual and normative *spaces* designed and accepted by humans. Technologies are *allowed* to vary their behavior within these spaces to the extent that humans consider an action to be meaningful rather than erroneous behavior. Autonomous agents operate within the boundaries of this acceptable behavior, constrained by stabilized rules and norms. The EO-1 thus performs its tasks autonomously within the constraints set by human developers and other human actors. Outside of these constraints the behavior is considered a flaw or failure. Given the right conditions - including the conviction of the human engineers that the agent would not endanger the safety of the spacecraft and that the spacecraft would deliver interesting images - the ASE was allowed to operate itself.

Are there limits to the form and configuration of spaces that humans create? How do new technologies affect these spaces? To what extent can agents be maneuvered into similar roles as human beings? Answers to these questions not only depend on the level to which technologies can be made to regulate themselves or to generate their own goals. It also depends on our conceptions of humans, technologies and the relations between them within particular contexts, as well as on the obduracy of the cultural and historically constituted asymmetries between humans and technologies.

4.5 LIMITS TO AUTONOMY

Collins and Kusch present the distinction between polimorphic and mimeomorphic action as an insurmountable asymmetry between humans and machines. Yet, their analysis also shows that the tendency to hold humans morally responsible, rather than machines, does not mean that responsibility is attributed in equal measures to all humans. Not every person is considered to be able or in the position to make moral decisions. As Collins and Kusch point out, in Taylorist-style organizations the need for independent thought or decision-making on lower levels of the hierarchy is reduced as much as possible. Moral responsibility, similar to autonomy, is a malleable concept that acquires meaning within particular contexts.[19] The extent to which a person can or should

[19] In considering the question whether computers can be held morally responsible, philosophical discussions of computing and moral responsibility have explored various aspects of moral responsibility, such as the different senses of responsibility and the

be considered an autonomous person who can be held morally responsible depends on a variety of factors, including cultural, economic and political interests. Acknowledging the context-dependent nature of 'moral responsibility' and 'autonomy', as well as their efficacy sheds a different light on the limits to the autonomy of artificial agents. It draws attention to the larger sociotechnical systems in which humans and technologies become connected.

To say that technological devices are generally not attributed moral responsibility, is not to say that technology does not play a role in moral action. Although humans delimit the space in which technologies perform, technologies in turn set conditions on the range of actions humans can perform, often in ways not anticipated in their design. Technological artifacts persuade, facilitate and enable particular human cognitive processes, actions or attitudes, while constraining, discouraging and inhibiting others (see Chapter 2 and 3). As the philosopher Peter Paul Verbeek points out, technological artifacts are "active mediators" that "actively co-shape people's being in the world; their perception and actions, experience and existence" (Verbeek, 2006, p. 364). They affect the decisions that humans make and how they make them, and thus shape moral actions (Akrich, 1992; Latour, 1992). A speed bump, for instance, can impose moral behavior on a human driving a car, by encouraging her to slow down (so as to not damage her car) and adhere to local traffic norms. It enforces particular morally desirable behavior, while it limits the possibility for the human driver to act otherwise.

The mediating role of technology makes the development of technology an inherently moral activity. Verbeek states that the constitutive role of technologies in action places technological mediation at the heart of ethics: "Ethics is about the question of how to act, and technologies appear to be able to give material answers to this question by inviting or even exacting specific forms of action when they are used" (p.377). This point of view shows that efforts to endow robots and computer systems with social and moral knowledge (i.e. 'building ethics into' these technological artifacts) to perform particular tasks do not constitute anything significantly different from what current developers of technology already do. Rather it merely explicates the role of designers and engineers as "doing ethics by other means" (p.369).

The active role of technological artifacts then raises the question of what values and moral knowledge are inscribed in artificial agents and

---

distributed character of moral responsibility (Coleman, 2005; Kuflik, 1999; Nissenbaum, 1994).

how these technologies shape actions and experiences. Researchers engaged in the development of 'moral' agents will run and indeed have run into the problem of which or whose ethics to 'build in', as the nature of ethical principles continues to be a topic of debate (Allan et al., 2000). In addition, the ambition to develop these agents can generate conflicts with prevailing norms and values, including the respect for the autonomy of persons. Autonomous (moral) agents can infringe on the autonomy of human actors. Verbeek notes that the notion of the 'moralization of technology', as once proposed by the philosopher Hans Achterhuis, has led to fierce critiques warning against enabling technocratic, Orwellian 'big brother' societies. The moralization of technology entails the explicit intention to develop technologies that enforce morally desirable behavior.[20] Thus, instead of only 'moralizing humans' by telling them not to use the shower too long or not to drink and drive, technologies can be delegated these tasks. A shower head can be developed that, upon reaching a threshold, automatically turns off the water, and an 'alcohol lock' can be installed in a car that requires the driver to pass a breathing test before she can start the car.[21]

Critics of 'moralizing technology', as Verbeek writes, have argued that it jettisons the democratic principles of our society and threatens human dignity. It deprives humans of their ability and rights to make deliberate decisions and to act voluntarily. In addition, critics have claimed that if humans are not acting in freedom their actions cannot be considered moral. These objections can be countered, as Verbeek notes, by pointing to the rules, norms, regulations and a host of technological artifacts that already set conditions for actions that humans are able or allowed to perform. Moreover, technological artifacts as active mediators affect the actions and experiences of humans, but they do not determine them. Nevertheless, the critiques underline the moral issues at stake in choosing and interpreting metaphors for the development of technolo-

[20] The emerging field of research on *Persuasive Technology* explicitly aims to develop technology to persuade humans to perform in 'desirable' ways (IJselsteijn et al., 2006).

[21] The anti-alcohol lock is already in use in a number of countries, including the USA, Canada, Sweden and the UK. In the Netherlands trials with the device are on the way. The British newspaper the Guardian stated: "The day when intelligent machines overrule dumb humans came closer yesterday with the British launch of a technology that refuses to allow a car to start if it detects even a whiff of alcohol on the breath of the driver"(Vidal, 2004). Nevertheless, the article then proceeds to describe the various ways in which the device is far from foolproof. Moreover, it describes the ingenuity of people in trying to circumvent the strict morality of the device, by, for instance, keeping an air pump handy in the car.

gies. Given the tension between the historically and culturally constituted asymmetries between humans and technologies on the one hand, and the active role of technologies on the other, how can we address the concerns about increasingly autonomous technologies?

The discussions in previous chapters showed that computer technology and its role or impact cannot be evaluated outside of the context of design, use or operation. In terms of the vocabulary of Actor Network Theory, the mediating role of technologies comes about in open-ended networks of heterogeneous human and non-human actors (Latour, 2005; Law & Hassard, 1999). Particular human/technology configurations are the product of the interactions between various interests, resources, knowledge systems, competences of human actors, deeply rooted beliefs and ideals, as well as the non-human nuts, bolts, electrons, and other material entities. The successful operation of the EO-1 spacecraft is an impressive example of technological progress. Characterizing this as the inevitable result of some quasi-evolutionary process of increasingly complex technology acquiring new competencies does not do justice to the considerable work that goes into the coordination and organization of the extensive networks in which the spacecraft is embedded. The finances contingent on the political climate and the willingness to invest in space technology research were as much a critical element of the process that led to the self-controlled spacecraft, as was the fine-tuning of the algorithm that prevents the spacecraft from bumping into other objects.[22]

It is when things go wrong that the complexity of the interdependencies of humans and technologies becomes visible. Tracing the sequence of events that led to the breakdown of a system usually leads investigators in many directions, including the nature of the surrounding social organization, the context of the design and developing process, and the organizational and cognitive effects of automation (Leveson & Turner,

[22] NASA's Demonstration of Autonomous Rendezvous Technology (DART) is a spacecraft that is designed to maneuver itself independently around a satellite without guidance from ground control. In April 2006 the spacecraft was successfully launched, but proceeded to bump in to the satellite and abort its mission prematurely, because it detected low onboard fuel levels. The mission cost around $110 million. NASA called it partly successful, because it showed that spacecraft can find a satellite in space without human interference. However, it did launch an investigation to find what "anomaly" caused the "mishap" (http://www.newscientistspace.com/ article/dn7303.html). In November 2005 the US Air Force launched its similar $82 million micro-satellite Experimental Satellite System-11 that did successfully conclude its mission (http://www.newscientistspace.com/ article/dn8260).

1993; Nissenbaum, 1994; Reason, 1990). To consider how computer technologies affect moral actions we therefore have to consider the sociotechnical environments in which human/technology relationships take shape (Johnson, 2001).

The complex, interdependencies between humans and computer technologies underscore that descriptive, abstract accounts of artificial agents provide insufficient means to consider how future technologies will change human/technology relationships. Yet, they also illustrate the limitations of liberal conceptions of autonomy and moral responsibility. The unanticipated interactions between humans and complex technologies make attributing responsibility in practice a problematic undertaking (see for example Coeckelbergh & Wackers, 2007). Pervasive, interconnected, computer technologies add a layer of complexity to this problem (Nissenbaum, 1994). The philosopher Jeroen van den Hoven for instance notes that the cognitive dependencies that new computer technologies create can limit the extent to which users can take or be ascribed responsibility (2002). These complex technologies, which are never fully free from errors, increasingly hide the theories, models and assumptions that they embody. They make it more difficult for users to assess the validity and relevance of the information that they present, while users are often under pressure to make choices based on this information. Moreover, effects like 'automation bias' (see Chapter 3) or a lack of alternative knowledge sources to validate beliefs can interfere with the users' ability to make appropriate decisions.

The limitations of traditional ethical vocabularies in thinking about the social and moral aspects of new information and communication technologies, have led some authors in the field of computer ethics to reconsider concepts like moral agency and responsibility (Allan et al., 2000; Floridi & Sanders, 2004; Stahl, 2004). For example, the philosophers Luciano Floridi and Jeff Sanders propose to extend the class of moral agents to include artificial agents. In light of the increasing complexity of computer technology and the prospect of progressively autonomous software agents, they argue that the anthropocentric bias in the concept of moral agency results from its association with responsibility. "The whole conceptual vocabulary of 'responsibility' and its cognate terms is completely soaked with anthropocentrism" (Floridi & Sanders, 2004). The authors contend that "the insurmountable difficulties for the traditional and now rather outdated view that a human can be found accountable for certain kinds of software and even hardware" demand a different approach (p. 372). They instead suggest that artificial agents

should be acknowledged as moral agents that can be held accountable, but not responsible. They draw a comparison between artificial agents and dogs as sources of moral actions. Dogs can be the cause of a morally charged action, like helping to save a person's life (think of search-and-rescue dogs) or damaging property. We can therefore identify them as moral agents even though we generally do not hold them morally responsible. A dog can be held accountable by correcting or punishing it. Correspondingly, although artificial agents cannot be held morally responsible, they can be held accountable for a moral action. From this it follows, so Floridi and Sanders claim, that if an artificial agent can be observed as being the cause of moral action then the agent is a moral agent.

Eliminating the anthropocentric bias in the concept of moral agent does not put an end to the tension between delegating control and the fear of losing it. The reconceptualization of moral agency, as proposed by Floridi and Sanders, exemplifies a decontexualized analysis in which analogies between humans and technologies are further elaborated. Their argument hinges on the assumption that artificial agents are "sufficiently informed, 'smart', autonomous and able to perform morally relevant actions independently of the human engineer who created them, causing 'artificial good' and 'artificial evil'" (p.367). A system is an agent when it is interactive, autonomous and adaptive. Autonomy, according to the authors "means that the agent is able to change state without direct response to interaction: it can perform internal transitions to change in its state. This property imbues an agent with a certain degree of complexity and independence from its environment" (p. 357). Their definition mirrors the descriptive conceptualization of autonomy as used by the agent researchers described above. It is already divorced from its normative connotation. They proceed by adopting an intentional-stance-like approach to formulate a conceptual framework in which moral agency can be cleansed of its anthropocentric bias. In other words, if an artificial agent can be conceived of as behaving like a *moral* agent at an appropriate "level of abstraction", i.e. it meets the criteria of being interactive, autonomous and adaptive, then it can be held accountable (p. 352). This conclusion however seems to follow from a circular argument. An action, Floridi and Sanders state, is "morally qualifiable if and only if it can cause moral good or evil" and "an agent is said to be a moral agent if and only if it is capable of morally qualifiable action"(p. 364). Yet, moral good or evil, in their account, result from the actions of an agent

like a dog or a human (a person killed by an earthquake is not the result of a moral action).

Floridi's and Sanders' decontextualized and ahistorical account of moral agents overlooks the normative and constitutive role of concepts like autonomy and moral responsibility. Daniel Dennett originally proposed the intentional stance as a strategy that can be usefully applied when it allows an observer to predict and thereby explain the actions of a complex entity by treating it as if were a rational agent (1996, p. 27). However, when considered from a broader sociotechnical perspective, the extent to which the intentional stance can successfully predict the behavior of an entity is not the only criterion for its usefulness. The appropriateness of this stance, from this point of view, is a contentious issue that is negotiated within social contexts. Whether we accept computer systems as accountable agents depends on normative and context-specific conceptions of humans and technologies. I might reluctantly hold my laptop 'accountable' for its malfunction, and if it runs on an 'open source' operating system I might even 'correct' it (note that this requires a particular kind of design practice and culture) (David, 2004). However, in a more critical domain, say crisis response, this approach does not suffice.

Holding artificial agents accountable for errors or harm postpones the question of who has to account for the conditions under which artificial agents are allowed to operate. This becomes apparent when Floridi and Sanders reveal what they perceive as the benefits that can be gained from their reconceptualization of moral agents (i.e. the source of a moral action). The advantage of holding artificial agents accountable, Floridi and Sanders claim, is that these agents can be dealt with directly rather than having to find their creator. Immoral agents can be modified or deleted. "We are less likely to assign responsibility at any cost, forced by the necessity to identify a human moral agent. We can liberate technological development of [artificial agents] from being bound by the standard limiting view" (p. 378). Yet, this leaves the issue of sorting out which party is responsible for dealing with immoral agents. If technologies continue to be designed for instrumental and functional purposes, then reconceiving accountability in this way does not make humans less morally responsible and ultimately accountable. Humans set the conditions under which technologies are allowed to operate. As long as we do not accept computers as capable of amending their behavior appropriately in contingent circumstances, we look for responsible

human actors to account for the consequences of technologies, and to take appropriate action.

The concerns about autonomous technologies can therefore not be diffused simply by redefining concepts on an abstract theoretical level or by building moral agents. The anthropocentric bias in moral responsibility is constituted by deeply rooted asymmetries. This does not mean that some final limit to the autonomy of agents exists; rather the limits are multiple and flexible depending on the conceptual frameworks and social contexts in which the concept acquires meaning. Levelling humans and technologies in terms of their autonomy amounts to overcoming the anthropocentric bias, but this requires more than extending the competences of technologies. It entails a change in context-specific conceptions of humans and technologies and how they relate. The malleability of the notions of autonomy and moral responsibility opens the door to alternative human/technology configurations in which moral responsibility does not serve as the same organizing principle. This, however, is not an inevitable consequence of technological progress, or of the development of self-regulating artificial agents.

In considering the choices in the development and use of artificial agents, we should act with caution when adopting the descriptive metaphorical concepts used by designers of technology. The suggestion that the increasing complexity and autonomy of technologies demand a disconnection of accountability from moral responsibility is misleading. Floridi's and Sander's suggested overarching strategy involves a (normative) choice over alternative ways of organizing the conditions under which technologies are developed and used in particular contexts. Conceiving of artificial agents as accountable, moral agents could broaden the space in which technologies can operate independently (and unpredictably). It would allow designers to focus more on exploring ways to 'manipulate' system behavior rather than on methodologies for micromanaging every aspect of the system. At the same time, holding artificial agents accountable (and not addressing the issue of moral responsibility), has consequences for the organization of sociotechnical systems. A preoccupation with building artificial agents as autonomous moral agents can distract from thinking about development and use practices that explicate the extent to which humans can be held responsible and accountable for the things they make or how these things affect other people.

Given the obduracy and significance of the prevailing anthropocentric bias in sociotechnical systems, characterizing the advent of increasingly

autonomous complex technologies as an inevitable consequence of technological progress is a misleading and potentially dangerous line of reasoning. It narrowly frames the debate about the possibilities of dealing with problems that the development of artificial agents promises to overcome. In her paper on accountability in computerized societies, Helen Nissenbaum warns that "the conditions under which computer systems are commonly developed and deployed, coupled with popular conceptions about the nature, capacities and limitations of computing" can create barriers to accountability (1997 , p. 43). The tendency to use the computer as a scapegoat to attribute blame for errors is one of them. Nissenbaum conceives of accountability as something very akin to answerability, which can be used as "a powerful tool for motivating better practices, and consequently more reliable and trustworthy systems" (ibid.). Accepting the explanation that it is 'the computer's fault', she argues, stands in the way of a "culture of accountability" that is aimed at maintaining clear lines of accountability. A culture of accountability is worth pursuing, because a developed sense of responsibility is a virtue to be encouraged, and it is valued because of its consequences for social welfare. Holding people accountable for the harms or risks caused by computer systems provides a strong motivation for minimizing them. Moreover, accountability can provide a starting point to assign just punishment. Nissenbaum's instrumental take on accountability shifts the focus to the sociotechnical system in which technologies are developed and used. It underscores that increasingly autonomous technologies are the result of choices in developing technologies, rather than an inevitable outcome.

4.6 CONCLUSION

The analogy between humans and artificial agents in terms of their autonomy has been a central concern in debates about the possibilities and risks of artificial agents. In this chapter I have taken a closer look at conflicting interpretations of autonomy in these debates, in order to consider their significance with regard to research on artificial agents. Concerns about the loss of control associated with the idea of increasingly autonomous agents are indicative of a tension between two distinct conceptions of autonomy. Although the two notions are rooted in different conceptual systems, their interplay in various discourses has profound effects on human/technology relationships. A further analysis of the tension illustrated that an obdurate asymmetry between humans

and technology remains that leaves the human as the ultimate morally responsible party. This asymmetry originates in the social and cultural structures in which human/technology relationships take shape. The anthropocentric bias in the notion of moral responsibility sets constraints on the space in which self-regulating technologies can operate, as it generates a requirement for predictability, transparency and controllability. At the same time, the active role of computer technologies sets constraints on the space in which humans can act autonomously. Autonomy as an abstraction tool in exploring new ways of thinking about the development of computer technology can therefore conflict with the value human autonomy.

What the foregoing discussion illustrates is that concerns about increasingly complex, opaque and independently-operating technologies cannot be resolved by decontextualized efforts to bridge the gap between humans and technologies. Autonomy can have many different interpretations that are instrumental in organizing the world. What autonomous agents can do and mean is, therefore, dependent on the conceptual frameworks and social contexts in which these technologies are conceptualized, developed and used. A preoccupation with an analysis of inherent properties of artificial agent and humans distracts us from considering which, why and how particular asymmetries between humans and technologies should be maintained or dissolved. Increasing complexity of technology does not necessarily require a further elaboration of the analogy between humans and artificial agents. A contextualized sociotechnical perspective that acknowledges the different interpretations and roles of metaphorical concepts highlights the multiplicity of human/technology configurations and the various ways in which these configurations take shape in different contexts.

The analysis in this chapter underlines the responsibility of individuals in constructing, advocating, pursuing, and accepting agent visions. Reconceptualizing concepts like moral agency and autonomy can have real material effects on the organization of sociotechnical systems. It influences the design of technologies, and shapes the practices in which these technologies are developed and used. A central concern in choosing and interpreting metaphors to develop technologies should therefore be an analysis of the changes that the elaboration of particular metaphors presupposes, and of the choices and values involved in configuring humans and technologies in sociotechnical systems. Such an investigation poses the questions of where to look for the preferable solution to a perceived problem. Can technology solve this or do we

prefer to look for the solution elsewhere? What conflicts does the proposed conceptualization of human/technology relationships generate?

The limits to the autonomy of agents are the subject of an important debate. This is not just because advances in technologies are making the formulation of robotic law's more pressing, but because when considered from a sociotechnical perspective it provides an arena in which to (re)addresses questions about what norms and ethical principles we value and why. A change in human/technology relationships that would make the analogy between human autonomy and autonomy of computers less contested would require a major shift not only in how we think about technologies, but also in our more fundamental beliefs about moral responsibility. It would mean a world in which moral responsibility and autonomy no longer serve as the same organizing principles as they do in modern liberal societies. This would perhaps be possible in worlds where robots are seen as capable of autonomous thought and moral reasoning, just like animals and other animate entities, or maybe in a world where humans are perceived of as similarly limited in their autonomy as compared to conventional technologies and can not be held morally responsible. Such a shift in thinking, however, is not an inevitable consequence of technological progress; rather it constitutes a normative choice about how we want to organize the world. The question, then, is not to whether (future) artificial agents can be conceived of as moral agents, but whether, to what extent and under what conditions it is desirable or moral to do so.

5. AGENTS OF CHANGE

In his classic play *Rossum's Universal Robots* (R.U.R), Karel Čapek introduced the term robot as referring to artificial, human-like entities for the first time.[1] In the play the young and the old Rossum both have their own motivations for creating human-like creatures. As Čapek himself explains,

> The odd inventor, Mr. Rossum (whose name in English signifies Mr. Intellect or Mr. Brain), is no more or less a typical representative of the scientific materialism of the last century. His desire to create an artificial man [. . .] is inspired by a foolish and obstinate wish to prove God unnecessary and absurd. Young Rossum is the young scientist, untroubled by metaphysical ideas; scientific experiment to him is the road to industrial production. He is not concerned to prove but to manufacture.
>
> *(1923, p. 79)*

The ambitions underlying research and development centered on artificial intelligent agents are sometimes reminiscent of the motivations of the young and the old Rossum. One important driving force behind the development of intelligent devices is the challenge of building a human-like machine, of building our own successors. Some might call it playing God. Others justify the objective by referring to the opportunities it provides to study and learn more about our own behavior, mind, cognition or humanness (Breazeal, 2002). At the same time, a large part of the research on intelligent technologies has been concerned with the more practical ambition to create new innovative applications that extend and amplify human intellect and support humans in their activities. Regardless of whether the ambition to develop intelligent technologies is 'foolish' or 'obstinate', the discussions in the previous chapters show that the two objectives are related, yet not necessarily compatible.

In this study I have looked at the relationships between visions of artificial agents and current technological developments to explore how we can meaningfully speak about the possibilities, limitations and risks of intelligent technologies. These visions cultivate the image of technologies

[1] In R.U.R. Čapek introduced the term robot as referring to artificial human-like entities for the first time. "My dear miss Glory, the Robots are not people. Mechanically they are more perfect than we are, they have enormously developed intelligence, but they have no soul" (Čapek, 1991, p. 9).

that move increasingly closer to humans in terms of their competences and qualities. Certain researchers, as we saw in Chapter 2, have argued for the development of artificial agents that would be less like tools and more like communicating entities that can cooperate with humans. According to these researchers, such electronic 'intelligent assistants', 'partners' or 'teammates' hold the promise of enabling more intuitive, effortless and thus better interactions between humans and technologies. The discussion in Chapter 3 turned the focus to narratives of adaptive, agent-based systems that would be capable of supporting and replacing humans in increasingly more cognitive tasks. These technologies would enable a more intimate, symbiotic, cognitive relationship between humans and technologies. Finally, I addressed visions of artificial agents endowed with moral decision-making abilities in Chapter 4. In light of the prospect of increasingly autonomous technologies, agent advocates have argued for the development of agents that move closer to humans in terms of their ability to reason about the moral and social aspects of their actions.

A central claim in this book is that visions of agent technologies, and intelligent technologies in general, should be understood as constituted by metaphorical concepts, which draw particular analogies between humans and technologies. To understand the meaning and role of these visions, with respect to technological development, we need to take a closer look at these metaphorical concepts. My discussions of agent visions have focused on a variety of meanings and roles of key metaphorical concepts that support definitions of artificial agents. Agent researchers have used interrelated metaphorical concepts like communication, adaptive systems, and autonomy to describe their notions of artificial agents. Defined as such, the notion of 'artificial agent' enables particular interpretations of what the envisioned agents should be capable of and how they should relate to humans. The discussions showed that these concepts serve important constitutive, descriptive and heuristic functions within scientific research and engineering practices. The concept of adaptive, agent-based systems, for example, provides a design metaphor to structure the understanding of complex computer systems (see Chapter 3). It also serves as a heuristic device to guide the search for new forms of computing and modeling human/technology relationships. The discussion in Chapter 2 demonstrated that the appeal to artificial agents provides an effective rhetorical tool to rally support for and draw attention to particular projects and causes. Through the associations of the concept with ideographs, such as "effortless", "more

time" and "freedom", as well as with other metaphors, such as "digital assistants" and "collaborating partner", narratives about future intelligent agents present seductive images of future societies.

The instrumental role of metaphorical concepts in agent research emphasizes the need for a contextualized analysis of how and why researchers use particular metaphors. Such an analysis offers an alternative perspective on the varying and problematic understandings of humans, technologies, and the relations between them. By turning our attention to the ambitions and goals underlying research efforts we can move beyond the endless debate of whether technologies can be human-like, in order to consider the more important issues that are at stake when discussing the consequences of the development of (intelligent) computer applications. In particular, it allows us to address in what sense and under what conditions technologies *should* or *should not* be considered to be human-like. In this final chapter I will discuss how a critical interrogation of metaphorically structured visions contributes to reflective research and development practices as well as to the broader debate on the social and ethical aspects of agent technologies, and intelligent technologies in general.

5.1 MOVING BEYOND THE GAP

The ambiguity of the term agent has been a central concern in agent research. In their analysis of the various proposed definitions of artificial agents Franklin and Graesser noted: "Workers involved in agent research have offered a variety of definitions, each hoping to explicate his or her use of the word 'agent'. [. . .] We suspect that each of them grew directly out of the set of examples of agents that the definer had in mind" (Franklin & Graesser, 1997, p. 21). To capture the variety of definitions Franklin and Graesser attempted to formulate a definition of autonomous agents that would capture the essences of being an agent, "knowing full well that it must fail around the edges". By subsequently adding restrictions to particular classes of agents they aimed to produce "a nomenclature of agents that could be used relatively unambiguously by researchers in the field, resulting in clearer communications" (p. 25). Their proposed taxonomy is one of a number of attempts to formulate an abstracted and generalized definition of intelligent agents (Luck et al., 2005; Weiß, 1999; Wooldridge, 2002). Although a unified or standardized definition of agents might facilitate communication, an exclusive focus on such abstract definitions produces a narrow view that inhibits a

nuanced debate about the possibilities, limitations and risks of agent-based technologies.

Decontextualized discussions on the nature of (artificial) agents can lead to what feminist theorist Donna Haraway describes as *fetishism* (1997). She criticizes modernist, humanist, scientific traditions for enabling and supporting fetishism of the products of their heterogeneous practices of technoscience. She characterizes this fetishism as a kind of reification that mistakes a non-literal substitute or *trope* for a nontropic, literal, real thing with intrinsic value. Tropes, including models and metaphors, "mark the non-literal quality of being and language"(p. 135). She explores the idea of fetishism through the 'gene' trope. The gene has gone from an abstraction or code used to describe a particular aspect of human life, to a concrete thing that generates value (take for example gene patenting). By taking a gene as the thing itself, rather than a code, the sociotechnical relations among humans and between humans and nonhumans that produced 'the gene' and its value disappear from view. "Fetishes obscure the constitutive tropic nature of themselves and of the world"(p. 136). The problem and danger of these fetishes, Haraway argues, is that they hide accountability and foreclose a debate about the interests, ambitions, goals, and normative frameworks from which they generate. Abstract conceptions of agents can have similar consequences.

The term agent is an abstraction, as many agent researchers are well aware of (Jennings & Wooldridge, 1998; Luck et al., 2005). Its meaning is contingent on the discourses in which it is produced. It can refer to humans as well as to organizations or other non-human entities that produce some effect. At the same time, as a central feature in discussions on free will and determination, it is intimately tied to the concept of human. In philosophical, sociological, economic and legal discourse the term has served in different ways to emphasize particular aspects of what it means to be human (e.g. to act, free will, to have agency) or to highlight certain features of human behavior (e.g. rational decision-making). In agent discourse, as the previous chapters demonstrate, the agent metaphor (or trope) is a pivotal point in a network of interrelated concepts and ideas, linking the various discourses.

A diffuse and wide range of projects based on different ambitions, goals and conceptions enlist the notion of artificial agents to explain or model various aspects of computer technologies. The agent community is constituted by an eclectic assortment of researchers not all of whom are concerned with approximating human intelligence or behavior as closely as possible. The majority of agent research and development

projects does not claim or aspire to develop computational systems with human-like intelligence. Instead these projects are primarily concerned with building innovative computer technologies, taking aspects of human behavior as their inspiration. Human skills and qualities are referred to merely to characterize the desired behavior of the system, not to explain or describe human behavior.

Hence, as I have argued in this book, in order to understand what envisioned agent technologies entail and to keep accountabilities in view, we have to move beyond discussions of essential qualities of agents. We have to look at why and how agent researchers develop their visions and in what contexts. A first step towards this end is to critically interrogate the metaphorical concepts that appear in agent discourse, informed by contextualized analyses of human/technology relationships. Conceiving of agent visions as metaphorically structured, reminds us that these visions embody particular conceptions, as they highlight and hide certain aspects of humans, technologies and the relations between them (Lakoff & Johnson, 1980). Using the concept of agent, researchers in the field of agent-based computing have drawn particular analogies between humans and technologies to describe how envisioned technologies should function in relation to humans, their environment and other technologies. Figuring humans and agents as two communicating social entities, for example, prioritizes social interaction as being the 'natural' mode of action for humans. Other commonly referenced features of human behavior are pushed to the background, such as the physical manipulation of the environment by an embodied individual (Dourish, 2001; Suchman, 1987, 2003). As Chapter 2 showed, the abstract descriptions of artificial agents and persuasive narratives of a natural evolution towards increasingly intelligent agents mask the assumptions, interests, ambitions and goals that underlie the particular conceptualizations of the envisioned human/technology relationships. Abstracted and deterministic accounts of changing human/technology relationships should therefore prompt us to ask what these visions represent, where they come from, what they do, and why they are so persuasive.

Throughout this book, I have challenged the assumption that the gap between the competences of humans and technologies should be bridged. I have contrasted agent visions that build on this assumption with theories and empirical studies from fields like STS, HCI, phenomenology and cognitive science. In particular, I have drawn on studies that explore the complementary relationships in which distinctions between humans and technologies play a constitutive role, and which stress

contextualized sociotechnical perspective on human/technology relationships. The purpose of this comparison has been to expose the particular interpretations of metaphorical concepts, rather than to argue against the continuous use of analogies between humans and technologies in computer science.

The various perspectives on human/technology relationships that I discussed remind us that the most optimal configuration of humans and technologies are not just the result of some intrinsic properties of human or technological devices. Shifting perspectives brought back into focus the multiple dimensions of human/technology relationships as well as their context-dependent and co-constitutive nature. Technological artifacts persuade, facilitate and enable particular human cognitive processes, actions or attitudes, while constraining, discouraging and inhibiting others. At the same time, the meaning and use of an artifact as well as our experience of it is shaped by our previous experiences, our background knowledge, our conceptual and normative systems as well as the circumstances under which we engage with the device (Bijker et al., 1987; Ihde, 2003; Norman, 1999). Technologies affect human action, but humans can appropriate, reconfigure or even reject these devices (Akrich, 1992).

The context-dependent, co-constituting processes between humans and technologies underline that agent visions represent normative accounts of what technologies *should do* in relation to humans, rather than *descriptions* of (future) human/technology relationships. These visions represent abstract and idealized conceptualizations of only a small subset of possible human/technology configurations. We should, therefore, be cautious of rhetoric that proposes an ultimate, optimal and necessary solution.

Nevertheless, as the discussion in this book show, dismissing imaginative narratives about the nature and the potential of future technologies as science fictional, ideological accounts of a handful of techno-enthusiastic scientists, or even as a marketing ploy is to disregard the significance of these visions as an element of sociotechnical systems. Metaphorically structured visions of technological change are inextricably linked to social, cultural and political environments. They are shaped by norms, values and culturally constituted systems of knowledge. In turn, these visions provide heuristic devices, abstraction tools and design metaphors that guide and shape research into new innovative technologies. They set issues on the research agenda, as they provide for collectively shared goals to aspire to and views on directions to take in

research and development practices. On a broader level visions of promising new technologies, as van Lente has shown, affect the 'strategic' decisions about technological development in policy circles and industry. Promises and expectations of technological development "are an integral part of the innovation process, playing a role in every phase of development" (van Lente, 1993, p. 8). The European financed ISTAG, for instance, presented 'ambient intelligence' as a wide ranging vision of how the information society will develop, and as a focus point for discussions around the requirements for ICT research in European funded programs (Ducatel et al, 2001).

In this book, I took a closer look at how visions of artificial agents relate to current research and development practices by focusing on the metaphorical concepts that support these visions. I analyzed the various meanings of the concepts of social interaction, adaptivity and autonomy in different contexts, drawing on relational and contextualized perspectives on the connections between humans and technologies. I examined the role of these concepts in shaping the understanding of agent technologies, what normative assumptions the different interpretations of these concepts reflect, and what conflicts these interpretations can generate. The presented analysis provides a basis for *a pragmatic approach to discussing the possibilities, limitations and risks of intelligent technologies*. Such an approach entails a reflection on the metaphorical concepts that constitute visions of intelligent technologies, informed by sociotechnical, contextualized analysis of human/technology relationships. The focus in the suggested approach is thus on empirically grounded analyses of the practical consequences of conceptualizing, researching, and developing innovative technologies. In the next section, I will address how this approach contributes to a broader researcher agenda.

5.2 A BROADER RESEARCH AGENDA

The preceding chapters showed that the consequences of the development and use of experimental computer technologies can only be theorized or empirically studied to a certain extent, as these technologies exist primarily within the confinements of laboratories. Moreover, they demonstrated that overarching technological solutions do not exist. Bridging the gap between humans and technologies can therefore not be considered an ultimate goal, but should be understood as one of a range of proposed models to structure the understanding of new computer technologies. Assessing these different options requires a contextualized

analysis of the metaphorical concepts used to describe the envisioned technologies in order to expose the *choices* and *conflicts* that the proposed technologies can generate in different contexts. Against the background of contextualized analyses of the connections between humans and technologies, we can then consider in what sense and under what conditions technologies should or should not move closer to humans.

Turning our attention to the metaphorical concepts used in agent discourse brings into view the choices involved in the conceptualization and development of the envisioned technologies. It provides a way to understand, evaluate and discuss the assumptions that underlie narratives of future technologies like artificial agents. These concepts, as the analyses in the previous chapters showed, are discursive elements that shape and are shaped by practices constituted by ideas, theories, methods, technologies, techniques, norms and values. By acknowledging that metaphorical concepts acquire meaning within particular discourses we can expose and examine the conflicts that result from the confrontation of the different meanings and roles of these concepts.

Choices

Visions of agents reflect different kinds of choices. To illustrate, I will briefly highlight three. First, they represent choices about the 'configuration' of human users (Woolgar, 1991). Agent visions build on definitions of the knowledge, skills and responsibilities that the envisioned human users should have, and of the actions that these technologies should enable or constrain. As we saw in Chapter 2, seductive visions of artificial agents as invisible personified agents, electronic assistants or teammates follow from a particular framing of the problem of human/technology interaction. In this framing, the 'effort' required to (learn to) operate conventional computers makes interacting with these devices 'burdensome' and 'unnatural'. Future computer technologies should therefore minimize this effort, reduce learning curves, and support intuitive interaction. Building on the metaphors of social interaction or human-like communication, agent advocates present agents with more human-like communicative and social competences as the most natural and optimal solution to the problem of human/computer interaction. However, the extent to which modeling the human/technology relationships after social interaction is a preferred solution is subject to debate. Intelligent agents that adjust and learn from human habits and preferences, present a trade-off (Schneiderman & Maes, 1997). Artificial agents that would be able to reduce the effort required to (learn to) work with complex technologies also imply the

creation of new dependencies that can lead to the reduction of control and responsibility on the human end. As we saw in Chapter 3, increased automation of decision-making tasks can prevent a human operator from effectively performing her tasks when it frustrates her understanding of how the system works.

Second, the choices that particular metaphorical concepts and their interpretations reflect pertain to the various possibilities in configuring humans and technologies in sociotechnical systems. The ambition to endow artificial agents with human-like skills, in order to bridge the gap between humans and technologies, conveys a commitment to a particular approach to organizing sociotechnical systems. Images of adaptive agents capable of operating independently in complex, dynamic environments suggest that cognitive competences, and thus decision-making tasks, can simply be transferred between humans and technologies. However, automation involves the deconstruction, reassembling, and reassigning of tasks and responsibilities across the chains that link humans and technologies. Hutchins' theory of Distributed Cognition, as discussed in Chapter 3, provided an alternative framework in which to consider what ensembles of humans and technologies can do and how they relate on a cognitive level (Hutchins, 1995). The ability of a distributed cognitive system (encompassing humans and technologies) to adapt is not a property that is reducible to the qualities of its individual components. This understanding of cognition highlights that in some cases leveling the competences between humans and technologies is undesirable. If the objective of the development of computers systems is to support human decision-making, then stable and predictable computer systems are often a preferable option.

Finally, metaphorically structured visions present a choice about design (and use) practices. For example, the suggestion that in light of the increasing complexity of our computerized world the development of progressively autonomous agents (and even moral agents) is an inevitable and necessary development is misleading. Insisting on conceiving of or constructing artificial agents as moral agents can obscure the responsibility of the creators and users of the technologies. The ambition to focus on the development of autonomous agents represents a normative choice about how to organize research and development practices (see Chapter 4). It preferences the development of increasingly complex technologies over the exploration of design practices that emphasize human responsibility and accountability (Nissenbaum, 1994). A broader sociotechnnical perspective on human-

technology relationships offers a wider range of possibilities to attend to this problem.

Conflicts

A second benefit of contextualized analyses of metaphorical concepts is that they shed light on various conflicts between different interpretations of these concepts. Conflicts can occur on multiple levels. I will address three kinds of conflicts to illustrate. First, metaphorical concepts can lead to conflicts in terms of the understanding of what technologies can do. The discussion in Chapter 3 showed that extracting concepts from the context in which they are meaningful can result in misplaced expectations of current and future technologies. The notion of adaptive, agent-based systems is meaningful within narrowly defined contexts. As a concept featuring in various discourses, the multiple connotations of adaptivity can cause confusion about the requirements and capabilities of computer technologies.

Second, the use of abstract, metaphorical concepts can result in inconsistencies within agent visions. Chapter 2 highlighted the often conflicting conceptions of humans and technologies underlying visions of artificial agents. In narratives about future computer technologies serving as assistants or team members, humans are cast as the ideal to aspire to, as they are capable of learning, adapting to and operating in complex environments. At the same, the human user is represented as an entity that is incapable of dealing with new situations and adapting to technologies. In turn, agent technologies are figured both as extensions of human activity as well as proactive, autonomous entities with a mind of their own. Such diverse and often conflicting conceptions yield inconsistent models of envisioned human/technology relationships.

Finally, the metaphorical concepts used to conceptualize future human/technology relationships can result in conflicts on a normative level. The models and theories inscribed in agent technologies can conflict with prevailing values within social contexts. Chapter 4 highlighted the tension that can arise from the confrontation of two different interpretations of autonomy rooted in different discourses. The suggestion that intelligent autonomous agents will or should be moral agents is problematic not only on an ontological level (see the final section of this chapter), but also on a practical level in the context of technological development practices. The concept of autonomous persons serves as an organizing principle in liberal democratic societies. It preferences particular human/technologies configurations, in which humans are the ultimate morally responsible party. The ambition to move humans and

technologies closer in terms of their autonomy can distract from organizing development and use practices such that they reflect such deeply rooted values and norms.

A pragmatic approach centered on exposing the choices and conflicts involved in constructing visions of future technologies contributes to broader debates on the promises and expectations of innovative agent technologies, as well as to more reflective research practices within the field of agent-based computing. An informed awareness of the inherent choices and conflicts that visions of artificial agents entail invites further discussion on the conditions under which the proposed technologies could be a part of particular sociotechnical organizations. Self-organizing MAS, for instance, provide an attractive framework for online auctions. However in the public sector domain they could easily lead to conflicts with prevailing values in this domain, such as transparency and accountability. Choices and conflicts provide a starting point to discuss the various dimensions of the conceptualization, development and (eventual) use of agent-based technologies. In this book I concentrated in particular on the normative, social and cognitive dimensions.

The choices and conflicts underline the responsibility of agent researchers in constructing, pursuing and advocating their visions. A critical interrogation of the metaphors that support conceptualizations of future technologies enable a more reflective approach to researching and developing innovative computer technologies. Such an interrogation contributes to what Agre has called a "critical technical practice":

> Instead of seeking foundations [a critical technical practice] would embrace the impossibility of foundations, guiding itself by a continually unfolding awareness of its own workings as a historically specific practice. It would make further inquiry into the practice of AI an integral part of the practice itself. It would accept that this reflexive inquiry places all of its concepts and methods at risk. And it would regard this risk positively, not as a threat to rationality but as a promise of a better way of doing things.
>
> *(1997, p. 23)*

Agre stresses that the object of critical reflection should not just be the computer systems, but also the process of technical work. Routinely rethinking premises, re-evaluating methods, and reconsidering concepts by examining their origins, enables AI researchers to comprehend and learn from the limitations of historically formed technical practices. For Agre a critical technical practice enables better ways of understanding

computation and how it can be used to learn about human nature, but it is equally relevant for research focused on exploring innovative computer technologies that would change human/technology relationships. Reflecting on the concepts that support particular understanding of humans/technology relationships should be an integral part of practices in this field of research. In addition, incorporating a contextualized and relational perspective on these relationships supports a flexible and more comprehensive approach to exploring new ways of connecting humans and technologies.

A contextualized and relational perspective places the focus on the connections between humans and technologies in sociotechnical systems, rather than on the abilities of isolated computer systems. This shift of focus suggests directions for further research on how to conceptualize and develop new kinds of computer technology. Verbeek, for instance, proposes that thinking in terms of technological mediation allows for alternative development and technological assessment practices that aim to establish a connection between the context of design and the context of use (2006).[2] It encourages designers to anticipate the future mediating role of a technological artifact and moral assessment of this role. Such an approach would benefit from a critical analysis of the metaphorical concepts that feature in these contexts, as different interpretations of metaphors can lead to different design decisions and different uses.

The consequences of developing intelligent technologies with regard to our conceptions of what it means to be human has been left unaddressed in the foregoing discussion of the proposed pragmatic approach. In order to discuss the possibilities, limitations and risks of intelligent technologies these consequences need to be considered. The ambition to develop increasingly intelligent technologies not only affects research and design practices, it challenges the traditional boundaries between humans and technologies. The contextualized relational perspective I have explored in this book sheds a different light on the debates about the blurring of boundaries between humans and technologies.

[2] Coeckelbergh and Wackers provide another perspective (Coeckelbergh & Wackers, 2007). They propose that a further investigation of the role of moral imagination in activities provides a basis for dealing with issues involving distributed responsibility.

> Late twentieth century machines have made thoroughly ambiguous the differences between natural and artificial, mind and body, self-developing and externally designed, and many other distinctions that used to apply to organisms and machines. Our machines are disturbingly lively, and we ourselves frighteningly inert.
>
> *(Haraway, 1991, p. 152)*

Blurring boundaries is not a thing of the future. In her book *The Second Self,* Sherry Turkle describes how computer devices, when they first started to appear on the market, challenged the boundaries between humans and technologies (Turkle, 2005). Through her studies in late 1970's and early 1980's of the interaction between individuals with the, at the time, novel devices, she demonstrated that these *evocative objects* seduced people to think of themselves in computational terms. Terms drawn from computer science discourses - think of 'processing', 'reprogramming' and 'debugging' - have become thoroughly interwoven with our vocabulary for talking about everyday psychology. We redefine ourselves through these technologies.

As Katherine Hayles points out in reference to Turkle "the co-constituting relation between humans and technologies has taken a new turn with the invention of the intelligent machine" (Hayles, 2005, p. 132). She remarks that "researchers with the greatest stake in developing these objects consistently use a rhetoric that first takes human behavior as the inspiration for machine design and then, in a reverse feedback loop, reinterprets human behavior in light of the machines" (p. 132). Since the invention of the computer, this feedback loop has consistently shaped the way we think about ourselves. The computational and information-processing paradigm has had a profound influence on academic disciplines concerned with human nature, most notably psychology and philosophy. Moreover, it has shaped common conceptions of what it means to be human (Edwards, 1996; Hayles, 1999). John Pickering sees the dynamics of this co-evolutionary relationship between humans and technologies as a reason to predict the advent of computational agents that will "make it difficult or unimportant to distinguish between technologised human agents and humanized technological artifacts" (1997, p. 45). However, this argument does not account for the construction of new boundaries and new categories.

Although humans mirror themselves in technology, they differentiate themselves from technology at the same time. The computer's interactivity and complexity positions it on the margin of known categories, such

as 'alive'. Turkle characterizes computers as objects "betwixt and between" psychology and the physical. Whereas traditionally we have been able to distinguish humans from other entities such as animals and machines on the basis of mental qualities, computers as devices that 'talk back' and that appear to reason rationally disrupt this scenario. They are not quite like animals or passive objects, but they are not like humans either. They are located in what the philosopher Ruud Hendriks calls a *vacuum of exemplars* (Hendriks, 2000, 2004). In his analysis of conceptions of a shared life of people with and without autism, Hendriks argues that we have historically developed a vocabulary of words, actions and norms to engage with either humans or things. Entities that are significantly different from ready-made exemplars resist categorization within the established conceptual framework constituted by dualist distinctions. Thus, autistic individuals end up in the grey area between humans and things. They present problems in regard to making them part of every day social life, because they do not fall within our expectations of how a socialized person behaves. In a similar sense, interactive computers that show increasingly more human-like behavior end up in a vacuum.

Computers encourage us to refine our categories, because of their position in the margins of categories. The computer, according to Turkle, upsets the traditional scenarios of what makes humans special. Drawing on Piaget's discovery of children as metaphysicians, Turkle studied how children develop theories to deal with and neutralize what seems threatening (2005). Things that are not understood or do not fit well in their conceptions of the world, she holds, are alarming and frightening for children, yet at the same time they are fascinating. She describes how children construct categories to deal with what can be construed as being alive or animate and refine these categories by distinguishing gradations of aliveness. As children grow older they adjust their conception of which entities are alive in confrontation with new examples. They develop a nuanced language with new categories and concepts, such as "sort of a life" that allows them to differentiate between entities and to construct conceptions of what makes humans special (p. 52). Turkle quotes a twelve year old programmer as saying:

> When there are computers as smart as people, the computer will do a lot of the jobs, but there will still be things for humans to do. They will run restaurants, taste the food, and they will be the ones who will love each other, have families and love each other. I guess they'll still be the only ones who go to church.
>
> *(p. 63)*

The refining of categories, as exhibited by the children that Turkle studied, provides some perspective on the persisting gap between humans and technologies. Fifty years of research in AI has not resulted in artificial intelligence indistinguishable from the exemplar it derived from. No doubt the state of the art has not fully matured yet, but another reason is that the exemplar has been redefined, reconceived, and transformed numerous times in light of the pursuit of artificial intelligence. AI has not explained intelligence; rather it has enriched and diversified our conceptions of intelligence, thinking and what it means to be human.

As long as the concept of intelligent artificial agent is a notion that is open to reconstruction and redefinition, it is unlikely that categorizing humans and computers systems in the same class of agents will not be contested. The development of technologies, like artificial agents, presents the possibility of technologies with new kinds of abilities that continue to occupy the vacuous space between known or established categories. They will encourage us to refine our conceptual systems. As children grow up with new 'interactive', 'learning' or 'autonomous' agent-based technologies, they will develop new categories and concepts, and amend our old ones to delineate humans from non-humans. In the co-evolution of humans and machines boundaries are shifted, rather than dissolved, and not always in the same way. The interesting question then is who or what is excluded as boundaries are shifted?

The shifting of boundaries is not a phenomenon that only occurs in those practices in which we are confronted with unfamiliar entities. Literature in STS and, in particular, in feminist studies of science has demonstrated that the boundaries between humans and non-humans are continuously negotiated within specific discourses. The concept of what it means to be human is constructed, malleable and often contested. It is the outcome of social, historical, and political processes and has much to do with power. Feminist theorist Monica Casper provides an illustration of the conflicts and exclusions generated by the different attributions of agency (Caspar, 1994) . Based on her ethnographic studies of technoscientific practices in fetal tissue research and fetal surgery, she explores the attributions of human, nonhuman and agency and how these attributions are differently grounded in various concrete practices. Thus, whereas in fetal surgery the fetus is rendered a (potential) person and patient, in fetal tissue research it is conceived of as dead organic human material to be used as disembodied tool. In addition, she analyzes how in the construction of the fetus as person (or patient) agency is attributed to the fetus at

the cost of the pregnant woman. Through the focus on the fetus as person, the pregnant mother is rendered invisible as human actor. The woman is constructed as a "technomaternal" environment that the patient inhabits, or as the "best heart-lung machine ever" in the words of one doctor. As the pregnant woman is rendered an environment, she is to some extent stripped of her agency and with that of her rights as an autonomous person. The rhetoric of intelligent assistants and adaptive decision-making systems conveys a similar tendency to emphasize the agency of these artificial agents, while figuring human users as abstract simplified entities.

In Chapter 2, I argued that figuring artificial agents as autonomous, adaptive and interactive entities can lead to problematic conceptions of the role of humans both in design and use. Computer scientist and virtual reality pioneer, Jaron Lanier is passionately opposed to artificial agents for this reason. "Agents make people redefine themselves into lesser beings. THAT is the monster problem" (Lanier, 1995, p. 67). According to Lanier in order for a user to treat a computer system as if it were an agent, she would have to reduce her own agency.

> Agents are the work of lazy programmers. Writing a good user-interface for a complicated task, like finding and filtering a ton of information, is much harder to do than making an intelligent agent. From a user's point of view, an agent is something you give slack to by making your mind mushy, while a user-interface is a tool that you use, and you can tell whether you are using a good tool or not.
>
> *(p. 68)*

Although Lanier highlights the limited agency attributed to humans that agent visions often imply, his comments also exemplify the same abstract and determinist view of human/technology relationships that make visions of agents problematic. Whether humans will indeed 'make their minds mushy' cannot be deduced from abstract technological visions. The dependencies between humans and technologies are shaped within local practices.

In choosing metaphors to conceptualize (future) human/technology relationships we should be sensitive to the effects of the boundaries we construct. They are an outcome, but a meaningful one. We draw boundaries to make sense of the world, but we do so at a cost. It is therefore important not to lose sight of the practices and the conceptual and normative frameworks that constitute the resulting distinctions and categories, such that we can question the biases and asymmetries that they harbor. In the words of Karen Barad: "Boundaries are not our

enemies; they are necessary for making meanings, but this does not make them innocent" (Barad, 1996, p. 187). She points out that constructed boundaries have real social and material consequences. By shifting perspectives and challenging boundaries and categories we are reminded of the concrete and local practices in which they originate. Barad's remark underlines that we cannot address questions about the boundaries between humans and technologies on the basis of a single ontology. We need an empirical philosophical approach that takes into account the particular sociotechnical practices in which these boundaries are constructed.[3]

Challenging boundaries allows us to question the conditions under which entities are socially configured as human, nonhuman and/or other. Suchman points out "if we take the human to be inseparable from specifically situated social and material relations, the questions shifts from 'will we be replicated?' to something more like 'in what sociomaterial arrangements are we differentially implicated, and with what political and economic consequences?'" (2003, p. 19). Restating the question in this way highlights the problematic nature of the claim that artificial agents will or should move increasing closer to humans. Future computational entities with which humans will engage in some form that is comparable to human social interaction, might well be a possibility. Nevertheless, narratives as presented by futurologists like Kurzweil about a progressive blurring of boundaries between humans and technologies masks the contentious shifting of boundaries. At what cost are agency, rights and other qualities attributed to technological artifacts or cyborg-like entities? Which entities will be conceived of as human, and what would constitute social interaction? These questions cannot be answered based on our current normative and conceptual frameworks.

Abstract conceptions of humans and technologies lead to discussions based on utopian dreams and technological doom scenarios that are of little relevance to debates about the social, political and ethical aspects of current efforts to develop intelligent technologies. The suggestion that a convergence of humans and technologies is the outcome of an evolutionary processes or that it is the most optimal configuration of humans and technologies leaves little room for the possibility of alternative configurations emerging between humans and technologies. When we take a closer look at the development of intelligent technologies through

[3] With Haraway, Barad argues for a focus on *situated knowledges*. Situated knowledges are accountable knowledges, as they do not divorce tropes (e.g. models and metaphors) from the sociotechnical practices in which they become meaningful (Haraway, 1991).

the lens provided by the concept of agents, it is by no means obvious that current developments will necessarily lead to computational entities with human-like intelligence and skills. In contrast to what imaginative narratives of futurologists suggest, it is difficult to identify a clear trend headed towards general purpose technologies that would make it *more* difficult to distinguish between humans and technologies.

In the beginning of this book, I stated that I wanted to move away from the preoccupation with the questions whether technologies can be like humans. I hope that this book has demonstrated that more interesting questions to ask are: *How, why and under which circumstances should we draw an analogy between humans and technologies, and when is it undesirable to do so*? My objective has been to show that it matters how we talk and think about technologies. Neither technologies nor metaphors are innocent or neutral. Reflecting on the ways in which we understand the gap between humans and technologies is therefore an essential element of any debate about the possibilities, limitations and risks of future intelligent technologies, for it is through conceptualizing this gap that we conceptualize and shape ourselves, our technologies and ultimately our society.

REFERENCES

Aarts, E., Collier, R., Loenen, E. v., & Ruyter, B. d. (Eds.). (2003). *Ambient Intelligence*. Berlin: Springer.

Aarts, E., Marzano, S., & Andrews, A. (2003). *The New Everyday: Views on Ambient Intelligence*. Rotterdam: 010 Publishers.

Adam, A. (1998). *Artificial Knowing: Gender and the Thinking Machine*. New York, NY: Routledge.

Agre, P., & Chapman, D. (1987). PENGI: An Implementation of a Theory of Activity. In *Proceedings of the Sixth National Conference on Artificial Intelligence (AAAI-87)* (pp. 268–272). Seattle, Washington.

Agre, P. E. (1997). *Computation and Human Experience*. Cambridge, UK: Cambridge University Press.

Akrich, M. (1992). The De-Scription of Technical Objects. In W. Bijker & J. Law (Eds.), *Shaping Technology/Building Society: Studies in Socio-Technical Change* (pp. 205-224). Cambridge, Massachusetts: The MIT press.

Allan, C., Varner, G., & Zinser, J. (2000). Prolegomena to any Future Artificial Moral Agent. *Journal of Experimental and Theoretical Artificial Intelligence, 12*, 251-261.

Alonso, E. (2002). AI and Agents: State of the Art. *AI Magazine, 23*(3), 25-29.

Arkin, R., & Moshinka, L. (2007). *Lethality of Autonomous Robots: An Ethical Stance*. Paper presented at the ICRA'07 IEEE International Conference on Robotics and Automation.

Asimov, I. (1950). *I, Robot*. New York, NY: Doubleday.

Barad, K. (1996). Meeting the Universe Halfway: Realism and Social Constructivism without Contradiction. In J. H. Nelson & J. Nelson (Eds.), *Feminism, Science and the Philosophy of Science* (pp. 161-194). Dordrecht: Kluwer Academic Publishers.

Bartneck, C. (2006). Reflection on Robotic Intelligence. *Proceedings of the CHI2006 Workshop on HCI and the Face, Montreal.*

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American, 284(5),* 34-43.

Biever, C. (2007). If You're Happy the Robot Knows it. *New Scientist Magazine, 2596* 30-31.

Bijker, W. E. (1995). *Of Bicycles, Bakelites, and Bulbs: Toward a Theory of Sociotechnical Change*. Cambridge, Massachusetts: The MIT Press.

Bijker, W. E., Hughes, T. P., & Pinch, T. (1987). *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology*. London, UK: The MIT Press.

Bijker, W. E., & Law, J. (1992). *Shaping Technology Building Society: Studies in Sociotechnical Change*. London, UK: The MIT Press.

Billings, C. E. (1997). *Issues Concerning Human-Centered Intelligent Systems: What's "human-centered" and what's the problem?* Paper presented at the NSF-HCS Workshop on Human-Centered Systems: Information, Interactivity, and Intelligence (invited talk). from http://www.ifp.uiuc.edu/nsfhcs/talks/billings.html.

Bishop, C. M. (1995). *Neural Network for Pattern Recognition*. Oxford, UK: Oxford University Press.

Bonabeau, E., Dorigo, M., & Theraulaz, G. (1999). *Swarm Intelligence: from Natural to Artificial Systems* (5th print ed.). New York, N.Y.: Oxford University Press.

Brachman, R. (2006). (AA)AI More than the Sum of its Parts. *AI Magazine, 27*(4), 19-34.

Bradshaw, J. M., Feltovich, P. J., Jung, H., Kulkarni, S., Taysom, W., & Uszok, A. (2004). Dimension of Adjustable Autonomy and Mixed-Initiative Interaction. In M. Nickles, M. Rovatsos & G. Weiss (Eds.), *Agents and Computational Autonomy: Potential, Risks, and Solutions* (pp. 17-39). Berlin; Heidelberg: Springer-Verlag.

Bradshaw, J. M., Sierhuis, M., Acquist, A., Feltovich, P., Hoffman, R., Jeffers, R., et al. (2003). Adjustable Autonomy and Human-Agent Teamwork in Practice: An Interim Report on Space Applications. In H. Hexmoor, C. Castelfranchi & R. Falcone (Eds.), *Agent Autonomy*. Boston; Dordrecht Kluwer Academic Publishers.

Breazeal, C. (2002). *Designing Sociable Robots*. Cambridge, Massachusetts: The MIT Press.

Breazeal, C., Brooks, A., Gray, J., Hoffman, G., Kidd, C., Lee, H., et al. (2004). Tutelage and Collaboration for Humanoid Robots. *International Journal of Humanoid Robotics, 1*(2), 314-358.

Breazeal, C., Buchsbaum, D., Gray, J., Gatenby, D., & Blumberg, B. (2005). Learning From and About Others: Towards Using Imitation to Bootstrap the Social Understanding of Others by Robots. *Artificial Life, 11*(1/2), 31-62.

Brooks, F. (1987). No Silver Bullet: Essence and Accidents of Software Engineering. *Computer, 20*(4), 10-19.

Brooks, R. (1990). Elephants Don't Play Chess. *Robotics and Autonomous Systems, 6*(1\2), 3-15.

Brooks, R. (1991). Intelligence Without Representation. *Artificial Intelligence, 47*, 139-160.

Brooks, R. A. (2002). *Flesh and Machines: How Robots Will Change Us*. New York, NY: Pantheon Books.

Bullock, S., & Cliff, D. (2004). *Complexity and Emergent Behaviour in ICT Systems*. Bristol, UK: HP Labs

Burghardt, P. (2004, May 2004). *COMBINED Systems: The Combined Systems Point of View*. Paper presented at the 1st International Workshop on Information Systems for Crisis Response and Management (ISCRAM), Brussels, Belgium.

Callon, M., & Latour, B. (1992). Don't Throw the Baby Out with the Bath School!: A Reply to Collins and Yearley In A. Pickering (Ed.), *Science as Practice and Culture*. Chicago: The University of Chicago Press.

Čapek, K. (1923). The Meaning of R.U.R. *Saturday Review July 21*, 79.

Čapek, T. B. (1991). *R.U.R. and The Insect Play*. Oxford, UK: Oxford University Press.

Carstensen, P. H., & Schmidt, K. (2003). Computer Supported Cooperative Work: New Challenges to Systems Design. In I. Kenji (Ed.), *Handbook of Human Factors/Ergonomics* (pp. 619-636). Tokyo: Asakura Publishing.

Caspar, M. J. (1994). Reframing and Grounding Nonhuman Agency: What Makes a Fetus an Agent. *American Behavioral Scientist, 37*(6), 839-855.

Castelfranchi, C. (2003). Formalising the Informal: Dynamic Social Order, Bottom-Up Social Control, and Spontaneous Normative Relations. *Journal of Applied Logic, 1*(1-2), 47-92.

Castelfranchi, C., & Falcone, R. (2003). From Automaticity to Autonomy: The Frontier of Artificial Agents. In H. Hexmoor, C. Castelfranchi & R. Falcone (Eds.), *Agent Autonomy:* (pp. 103-137). Boston; Dordrecht: Kluwer Academic Publishers.

Castelfranchi, C., & Falcone, R. (2004). Founding Autonomy: The Dialectics Between (Social) Environment and Agent's Architecture and Powers. *Lecture notes in computer science, 2969*, 40-54.

Ceruzzi, P. (2003). *A History of Modern Computing* (2nd ed.). Massachusetts: The MIT Press.

Chalmers, D. (1992). Subsymbolic Computation and the Chinese Room. In J. Dinsmore (Ed.), *Symbolic and Connectionist Paradigms.* . Hillsdale: L. Erlbaum Associates.

Chien, S., Sherwood, R., Tran, D., Cichy, B., Rabideau, G., Castano, R., et al. (2005, July 2005). *Lessons Learned from Autonomous Sciencecraft Experiment.* Paper presented at the Autonomous Agents and Multi-Agent Systems Conference (2005), Utrecht, Netherlands.

Christman, J. (2003a). Autonomy in Moral and Political Philosophy. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*.

Christman, J. (2003b). Autonomy in Moral and Political Philosophy [Electronic Version]. *The Stanford Encyclopedia of Philosophy* from http://plato.stanford.edu/archives/fall2003/entries/autonomy-moral/.

Christman, J., & Anderson, J. (Eds.). (2005). *Autonomy and the Challenges to Liberalism: New Essays*. Cambridge, UK: Cambridge University Press.

Christofferson, K., & Woods, D. D. (2002). How to Make Automated Systems Team Players. In E. Salas (Ed.), *Advances in Human Performance and Cognitive Engineering Research,* (Vol. 2). Greenwhich, CT: JAI Press, Elsevier.

Cichy, B., Chien, S., Schaffer, S., Tran, D., Rabideau, G., & Sherwood, R. (2004). *Validating the Autonomous EO-1 Science Agent.* Paper presented at the International Workshop on Planning and Scheduling for Space (IWPSS 2004), Darmstadt, Germany.

Clancey, W. J. (1997). *Situated Cognition: on Human Knowledge and Computer Representation.* Cambridge, UK: Cambridge University press.

Clark, A. (2003). *Natural Born Cyborgs: Mind, Technologies, and the Future of Human Intelligence*. New York: Oxford University Press.

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

Coeckelbergh, M., & Wackers, G. (2007). Imagination, Distributed Responsibility and Vulnerable Technological Systems: the Case of Snorre A. *Science and Engineering Ethics, 13*(2), 235-248.

Cohen, P., & Levesque, H. (1990). Persistence, Intention, and Commitment. In P. Cohen, J. Morgan & M. E. Pollack (Eds.), *Intentions in Communication*. Cambridge, Massachusetts: The MIT press.

Coleman, K. G. (2005). Computing and Moral Responsibility [Electronic Version]. *The Stanford Encyclopedia of Philosophy (Spring Edition 2005)* from http://plato.stanford.edu/archives/spr2005/entries/computing-responsibility/.

Collins, H. (1990). *Artificial Experts: Social Knowledge and Intelligent Machines*. London, UK: The MIT Press.

Collins, H., & Kusch, M. (1998). *The Shape of Actions: What Humans and Machines Can Do.* Cambridge, Massachusetts: The MIT Press.

Cummings, M. L. (2004). *Automation Bias in Intelligent Time Critical Decision Support Systems.* Paper presented at the AIAA 1st Intelligent Systems Technical Conference, Chicago.

Dautenhahn, K. (2002). *Socially Intelligent Agents: Creating Relationships with Computers and Robots*. Dordrecht: Kluwer Academic Publishers.

David, S. (2004). Opening the Sources of Accountability [Electronic Version]. *First Monday* 9, 1-27 from http://firstmonday.org/issues/issue9_11/david/.

de Oude, P., Ottens, B., & Pavlin, G. (2005). Information Fusion with Distributed Probabilistic Networks. In M. H. Hamza (Ed.), *Proceedings of the International Conference on Artificial Intelligence and Applications* (pp. 195-201).

de Wilde, R. (2000). *De Voorspellers: een kritiek op de toekomstindustrie*. Amsterdam, The Netherlands: De Balie.

Decker, K., Sycara, K., & Williamson, M. (1997). Middle Agents for the Internet?, *Proceedings of the 15th International Joint Conference on Artificial Intelligence*. Nagoya, Japan.

Dennett, D. C. (1993). *The Intentional Stance*. London, UK: The MIT Press.

Dennett, D. C. (1996). *Kinds of Minds*. London, UK: Weidenfeld & Nicolson.

Docampo Rama, M. (2001). *Technology Generations Handling Complex User Interfaces*. Ph.D. dissertation, Technische Universiteit Eindhoven. Eindhoven, The Netherlands: Universitaire Drukkerij.

Dorais, G. A., Bonasso, R. P., Kortenkamp, D., Pell, B., & Schreckenghost, D. (1998). Adjustable Autonomy for Human-Centered Autonomous Systems on Mars. In *Proceedings of the First International Mars Society Convention*. Boulder, CO,.

Dourish, P. (2001). *Where the Action is: The Foundations of Embodied Interaction*. Cambridge, Massachusetts: The MIT Press.

Dreyfus, H. L. (1992). *What Computers still Can't Do: A Critique of Artificial Reason* (2e ed.). Cambridge, Massachusetts: The MIT Press.

Ducatel, K., Bogdanowicz, M., Scapolo, F., Leijten, J., & Burgelman, J.-C. (2001). *Scenarios for Ambient Intelligence in 2010*. Luxembourg: Office for Official Publications of the European Communities.

Edwards, P. N. (1994). From "Impact" to Social Process: Computers in Society and Culture. In S. J. e. al. (Ed.), *Handbook of Science and Technology Studies* (Revised paperback edition ed., pp. 257-285). Thousand Oaks, CA, USA SAGE Publications.

Edwards, P. N. (1995). Cyberpunks in Cyberspace: the Politics of Subjectivity in the Computer Age. In *The Cultures of Computing*. Oxford, UK: Blackwell Publishers.

Edwards, P. N. (1996). *The Closed World: Computers and the Politics of Discourse in Cold War America*. Cambridge, Massachusetts: The MIT Press.

Elio, R., & Petrinjak, A. (2005). Normative Communication Models for Agent. *Autonomous Agents and Multi-Agent Systems, 11*(3), 273-305.

Encyclopædia Britannica. (2007). Automaton. from Encyclopædia Britannica Online: http://www.britannica.com/EBchecked/topic/44951/automaton (last accessed 14 November 2008).

Endsley, M. R. (1996). Automation and Situation Awareness. In R. Parasuraman & M. Mouloua (Eds.), *Automation and Human Performance: Theory and Applications* (pp. 163-181). Mahwah, Nj: Lawrence, Erlbaum.

Endsley, M. R. (2001). Designing for Situation Awareness in Complex Systems. In *Proceedings of the Second International Workshop on Symbiosis of Humans, Artifacts and Environment*. Kyoto, Japan.

Englebart, D. (1963). *Augmenting Human Intellect: A Conceptual Framework*: Stanford Research Institute, Menlo Park, CA.

Erickson, T. (1990). Working with Interface Metaphors. In B. Laurel (Ed.), *The Art of Human Computer Interface Design* (pp. 65-75). New York, NY: Addison-Wesley.

Erickson, T. (1997). Designing Agents as if People Mattered. In J. Bradshaw (Ed.), *Intelligent Agents*. Menlo Park, CA: AAAI Press.

Erickson, T. (2002). Some Problems with the Notion of Context-Aware Computing. *Communications of the ACM, 45*(2), 102-104.

Falcone, R., & Castelfranchi, C. (2001). The Human in the Loop of a Delegated Agent: The Theory of Adjustable Social Autonomy. *IEEE Transactions on Systems, Man and Cybernetics, 31*(5), 406-418.

Feinberg, J. (1989). Autonomy. In J. Christman (Ed.), *The Inner Citadel* (pp. 27-53): Oxford University Press.

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

Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A Survey of Socially Interactive Robots. *Robotics and Autonomous Systems, 42* 143-166.

Ford, K. M., & Hayes, P. J. (1998). On Computational Wings: Rethinking the Goals of Artificial Intelligence. *Scientific American, 9*(4), 78-83.

Forsyth, D. (1993). Engineering Knowledge: the Construction of Knowledge in Artificial Intelligence. *Social Studies of Science, 23*, 445-477.

Franchi, S., & Güzeldere, G. (Eds.). (2005). *Mechanical Bodies, Computational Minds: Artificial Intelligence from Automata to Cyborgs*. Cambridge, Massachusetts: The MIT Press.

Franklin, S., & Graesser, A. (1997). Is it an Agent, or Just a Program?: A Taxonomy for Autonomous Agents. *Lecture Notes in Computer Science, 1193*, 21-36.

Friedman, B., & Nissenbaum, H. (1997). Software Agents and User Autonomy. In W. Lewis Johnson (Ed.), *Proceedings of the First International Conference on Autonomous Agents* (pp. 466 - 469). New York, NY: ACM.

Gershenson, C., & Heylighen, F. (2003). Methodologies and Applications - When Can We Call a System Self-Organizing? *Lecture Notes in Computer Science, 2801*, 606-614.

Gibson, J. J. (1986). *The Ecological Approach to Visual Perception*. Hillsdale: Erlbaum.

Giere, R. N. (2002). Discussion Note: Distributed Cognition in Epistemic Cultures. *Philosophy of Science, 69*, 637–644.

Graesser, A., Person, N., Harter, D., & Group, T. T. R. (2001). Teaching Tactics and Dialog in AutoTutor. *International Journal of Artificial Intelligence in Education, 12*, 257-279.

Graesser, A., VanLehn, K., Rosé, C., Jordan, P., & Harter, D. (2001). Intelligent Tutoring Systems with Conversational Dialogue. *AI Magazine* (Winter).

Graubard, e. S. (1988). *The Artificial Intelligence Debate: False Starts, Real Foundations*. Cambridge Massachusetts: The MIT Press.

Habershon, E., & Woods, R. (2006, June 18, 2006). No Sex Please, Robot, Just Clean the Floor. *The Sunday Times*.

Hackett, E. J., Amsterdamska, O., Lynch, M., & Wajcman, J. (Eds.). (2008). *The Handbook of Science and Technology Studies, Third Edition* (3 ed.). Cambridge, Massachusetts: The MIT Press.

Haraway, D. (1991). *Simians, Cyborgs, and Women*. New York: Routledge.

Haraway, D. (1997). *Modest_Witness@Second_Millennium.FemaleMan©_Meets_OncoMouse-™*. New York, NY: Routledge.

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

Hayles, N. K. (2005). Computing the Human. *Theory, Culture & Society : Explorations in Critical Social Science, 22*(1), 131-152.

Hendriks, R. (2000). *Autistisch Gezelschap: Een Empirisch-filosofisch Onderzoek naar het Gezamelijk Bestaan van Autistische en Niet-autistische Personen*. Lisse: Swets and Zeitlinger.

Hendriks, R. (2004). Een Autistisch Verbeeldingstekort: Enkele Lessen uit de Psychologie en Meteorologie. *Algemeen Nederlands Tijdschrift voor Wijsbegeerte, 96*(1), 66-80.

Hill, T. (1989). The Kantian Conception of Autonomy. In J. Christman (Ed.), *The Inner Citadel : Essays on Individual Autonomy* (pp. 91-105). New York, N.Y.: Oxford University Press.

Hoffman, R. R., Feltovich, P. J., Ford, K. M., & Woods, D. D. (2002). A Rose by Any Other Name...Would Probably be Given an Acronym. *IEEE Intelligent Systems, 17*(4), 72-80.

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

Holland, J. H. (1996). *Hidden Order: How Adaptation Builds Complexity*. New York: Helix Books.

Hollnagel, E. (Ed.). (2003). *Handbook of Cognitive Task Design*. Mahwah, New Jersey: Lauwerence Erlbaum Associates.

Hutchins, E. (1995). *Cognition in the Wild*. Cambridge Massachusetts: MIT Press.

Hutchins, E., Keller, J. D., Bazerman, C., & Latour, B. (1996). Cognition in the Wild. *Mind, Culture, and Activity : an International Journal, 3*(1), 46-68 (23).

Hutchins, E., & Klausen, T. (1996). Distributed Cognition in an Airline Cockpit. In Y. Engeström & D. Middleton (Eds.), *Cognition and Communication at Work* (pp. 15-34). New York: Cambridge University Press.

Ihde, D. (2003). A Phenomenology of Technics. In R. C. Scharff & V. Dusek (Eds.), *Philosophy of Technology : the Technological Condition : an Anthology* (pp. 507-529). Malden, MA: Blackwell Publ.

IJselsteijn, W., de Korte, Y., Midden, C., Eggen, B., & Hoven, E. (Eds.). (2006). *Persuasive Technology*. Berlin: Springer-Verlag.

Ishiguro, H., & Nishio, S. (2007). Building Artificial Humans to Understand Humans *Journal of Artificial Organs, 10*(3), 133-142.

Ishii, H., & Ullmer, B. (1997). Tangible Bits: Towards Seamless Interface between People, Bits, and Atoms. In *CHI '97: Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 234-241). New York, NY: ACM Press.

ISTAG. (2003). *Ambient Intelligence: from vision to reality*.

Jennings, N. R., Sycara, K., & Wooldridge, M. (1998). A Roadmap of Agent Research and Development. *Autonomous Agents and Multi-Agent Systems, 1*, 275–306.

Jennings, N. R., & Wooldridge, M. J. (1998). *Agent Technology: Foundations, Applications, and Markets*. Berlin: Springer.

Jensen, F. V. (2001). *Bayesian Networks and Decision Graphs* (3rd print. ed.). New York, N.Y., etc.: Springer.

Johnson, D. G. (2001). *Computer Ethics* (3 ed.). Upper Saddle River, New Jersey: Prentice Hall.

Johnson, D. G. (2006). Computer Systems: Moral Entities but not Moral Agents. *Ethics and Information Technology, 8*, 195-204.

Johnson, S. (1997). *Interface Culture*. New York, NY: Basic Books.

Joy, B. (2000). Why the Future Doesn't Need Us - Our Most Powerful 21st-Century Technologies - Robotics, Genetic Engineering, and Nanotech - are Threatening to make Humans an Endangered Species. *Wired, 8*(4), 238-264.

Kay, A. (1972). A Personal Computer for Children of All Ages. In *Proceedings of the ACM National Conference*. Boston.

Kay, A. (1990). User interface: A Personal View. In B. Laurel (Ed.), *The Art of Human Computer Interface Design* (pp. 191-207). New York, NY: Addison-Wesley.

Kitano, H., Tadokoro, S., Noda, I., Matsubara, H., Takahashi, T., Shinjoh, A., et al. (1999). Robocup Rescue: Search and Rescue in Large-Scale Disasters as a Domain for Autonomous Agents Research. *IEEE Systems, Man and Cybernetics, VI*, 739–743.

Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. A. (1997). Intelligent Tutoring Goes To School in the Big City. *International Journal of Artificial Intelligence in Education, 8*, 30-43.

Kostakos, V., O'Neill, E., Little, L., & Sillence, E. (2005). The Social Implications of Emerging Technologies. *Interacting with Computers, 17*(5).

Kuflik, A. (1999). Computers in Control: Rational Transfer of Authority or Irresponsible Abdication of Authority? *Ethics and Information Technology, 1*, 173-184.

Lakoff, G., & Johnson, M. (1980). *Metaphors We Live By*. Chicago, US: University of Chicago Press.

Lanier, J. (1995). Agent of Alienation. *Interactions 2*, 66-72.

Latour, B. (1987). *Science in Action: How to Follow Scientists and Engineers Through Society*. Cambridge, Massachusetts: Harvard University Press.

Latour, B. (1992). Where are the Missing Masses? The Sociology of a Few Mundane Artefacts. In W. Bijker & J. Law (Eds.), *Shaping Technology/Building Society: Studies in Socio-Technical Change* (pp. 225-258). Cambridge, Massachusetss: The MIT press.

Latour, B. (2005). *Reassembling the Social : an Introduction to Actor-Network Theory*. Oxford, UK: Oxford University Press.

Lauwaert, M. (2007). *Changing Practices, Shifting Sites*. Unpublished PhD thesis, University of Maastricht.

Law, J. (1999). After ANT: Complexity, Naming and Topology. In J. Law & J. Hassard (Eds.), *Actor Network Theory and After* (pp. 1-14). Oxford: Blackwell.

Law, J., & Hassard, J. (1999). *Actor network theory and after*. Oxford: Blackwell.

Leveson, N. G., & Turner, C. S. (1993). An investigation of the Therac-25 accidents. *IEEE Computer, 7*(26), 18-41.

Lewin, R. (1993). *Complexity: Life on the Edge of Chaos*. London: Phoenix Orion Books.

Licklider, J. (1960). Man-Computer Symbiosis. *IRE Transactions on Human Factors in Electronics, HFE-1*, 4-11.

Lieberman, H., & Selker, T. (2000). Out of Context: Computer Systems that Adapt to, and Learn from, Context. *IBM systems journal, 39*(3/4), 617-633.

Luck, M., McBurney, P., & Preist, C. (2004). A Manifesto for Agent Technology : Towards Next Generation Computing. *Autonomous Agents and Multi-Agent Systems, 9*(3), 203-252 (250).

Luck, M., MCBurney, P., Shehory, O., & Willmot, S. (2005). *Agent Technology Roadmap: a Roadmap for Agent Based Computing*: the European Coordination Action for Agent-Based Computing.

Luck, M., Munroe, S., & d'Inverno, M. (2003). Autonomy: Variable and Generative. In H. Hexmoor, C. Castelfranchi & R. Falcone (Eds.), *Agent Autonomy* (pp. 9-22): Kluwer.

Maasen, S., & Weingart, P. (1995). Metaphor-Messengers of Meaning: A Contribution to an Evolutionary Sociology of Science. *Science Communications, 17*(1), 9-31.

Maasen, S., & Weingart, P. (2000). *Metaphors and the Dynamics of Knowledge*. London, UK: Routledge.

Mackenzie, D., & Wajcman, J. (1999). *The Social Shaping of Technology* (2nd ed.). Buckingham, UK: Open University Press.

Maes, P. (1994a). Agents that Reduce Work and Information Overload. *Communications of the Association for Computing Machinery, 37*(7), 31-41.

Maes, P. (1994b). Modeling Adaptive Autonomous Agents. *Artificial Life, 1*(1-2), 135-162.

Mambrey, P., & Tepper, A. (1996). Metaphors and System Design. In P. Hoschka (Ed.), *Computers as Assistants: A New Generation of Support Systems* (pp. 269-280). Mahwah, NJ.: Lawrence Erlbaum Associates.

Maris, M., & Pavlin, G. (2006). Distributed Perception Networks for Crisis Management. In *Proceedings of the Third International Conference on Information Systems for Crisis Response and Management (ISCRAM 2006),* (pp. 376-381). New Jersey, USA.

Markopoulos, P., Ruyter, B. d., Privender, S., & Breemen, A. v. (2005). Special Section : Ambient intelligence - Social Intelligence in Home Dialogue Systems. *Interactions : New Visions of Human-Computer Interaction, 12*(4), 37-43.

Maturana, H. R., Varela, F. J., & Paolucci, R. (1988). *The Tree of Knowledge: the Biological Roots of Human Understanding*. Boston, MA: New Science Library.

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (1955). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence.

McCorduck, P. (1979). *Machines who Think*. San Francisco, US: W.H. Freeman and Company.

Michalewicz, Z. (1999). *Genetic Algorithms + Data Structures = Evolutionary Programs*. Berlin: Springer-Verlag.

Minsky, M. (1988). *The Society of Mind*. New York: Simon and Schuster.

Mitchell, T. M. (1997). *Machine learning*. New York etc.: WCB/McGraw-Hill.

Moravec, H. (1990). *Mind Children: The Future of Robot and Human Intelligence*: Harvard University Press.

NASA. (2006, 24 January 2006). Autonomous Sciencecraft Experiment. Retrieved January 31, 2006, from http://ase.jpl.nasa.gov/

Negroponte, N. (1995). *Being Digital*. New York: Knopf.

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

Nickles, M., Rovatsos, M., & Weiss, G. (Eds.). (2004). *Agents and Computational Autonomy: Potential, Risks, and Solutions*. Berlin; Heidelberg: Springer-Verlag.

Nissenbaum, H. (1994). Computing and Accountability. *Communications of the Association for Computing Machinery, 37*(1), 72-80.

Nissenbaum, H. (1997). Accountability in a Computerized Society. In B. Friedman (Ed.), *Human Values and the Design of Computer Technology Book Contents* (pp. 41 - 64). Cambridge: Cambridge University Press.

Norman, D. (1993). *Things that Make Us Smart: Defending Human Attributes in the Age of the Machine*. New York, NY: Perseus Books.

Norman, D. A. (1999). *The invisible computer: Why good products fail, the personal computer is complex , and information appliances are the solution*. Cambridge, Massachusetts, London, England: The MIT Press.

Nourbakhsh, I., Sycara, K., Kroes, M., Yong, M., Lewis, M., & Burion, S. (2005). Human-Robot Teaming for Search and Rescue. *Pervasive Computing*(January), 72-78.

Nunnink, J., & Pavlin, G. (2006). Towards Robust State Estimation with Bayesian Networks: A New Perspective on Belief Propagation. In T. Arai, R. Pfeifer, T. Balch & H. Yokoi (Eds.), *Proceedings of the 9th Conference on Intelligent Autonomous Systems (IAS-9)* (pp. 722 -731).

Nwana, H., S. (1996). Software Agents: an Overview. *The Knowledge Engineering Review, 11*(3), 205-244 (240).

Nwana, H. S., & Ndumu, D. T. (1999). A Perspective on Software Agents Research. *The Knowledge Engineering Review, 14*(2), 1-18.

Oomes, A. H. J. (2004). Organization Awareness in Crisis Management In *Proceedings of the International Workshop on Information Systems for Crisis Response and Management (ISCRAM2004)*. Brussels.

Oudshoorn, N., & Pinch, T. (Eds.). (2003). *How Users Matter: The Co-Construction of Users and Technology*. Cambridge, Massachusetts: The MIT Press.

Parasuraman, R., & Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors : the Journal of the Human Factors Society, 39*(2), 230-253 (224).

Pavlin, G., de Oude, P., & Nunnink, J. (2005). A MAS Approach to Fusion of Heterogeneous Information. In *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence* (pp. 802 - 804): IEEE Computer Society.

Pavlin, G., Maris, M., & Nunnink, J. (2004). An Agent Based Approach to Distributed Data and Information Fusion. In *Proceedings of the IEEE/WIC/ACM International Joint Conference on Intelligent Agent Technology (IAT 2004)* (pp. 466 - 470). Beijing, China.

Perrow, C. B. (1999). *Normal accidents : living with high-risk technologies*. Princeton, NJ etc.: Princeton University Press.

Peters, P. (2006). *Time, Innovation and Mobilities: Travel in Technological Cultures*. New York, USA: Routledge Taylor and Francis Group.

Petrosky, H. (1992). *The Evolution of Useful Things*. New York, NY: Vintage Books.

Picard, R. (1997). *Affective Computing*. London, UK: The MIT Press.

Pickering, J. (1997). Agents and Artefacts. *Social Analysis : Journal of Cultural and Social Practice, 41*(1), 46-63.

Punie, Y. (2003). *A social and technological view on Ambient Intelligence in Everyday Life: What bends the trend?,* : European Media, Technology and Everyday Life Research Network.

Reason, J. (1990). *Human error*. Cambridge etc.: Cambridge University Press.

Reeves, B., & Nass, C. (1996). *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge: Cambridge University Press.

Rich, E., & Knight, K. (1991). *Artificial Intelligence* (2nd ed.). London: McGraw-Hill.

Riley, V. (1996). What Avionics Engineers Should Know About Pilots and Automation. *Aerospace and Electronic Systems Magazine, IEEE, 11*(5), 3-8.

Rocha, L. M. (2001). Adaptive Recommendation and Open-ended Semiosis. *Kybernetes, 30*(5-6).

Sack, W. (1997). Artificial Human Nature. *Design Issues, 13*, 55-64.

Sarter, N. B., Woods, D. D., & Billings, C. E. (1997). Automation Surprises. In G. Salvendy (Ed.), *Handbook of Human Factors and Ergonomics* (2 ed., pp. 1926-1943).

Scerri, P., Pynadath, D., & Tambe, M. (2002). Towards Adjustable Autonomy for the Real-world. *Journal of AI Research 17*(171-228).

Schneiderman, B. (1983). Direct Manipulation: A Step Beyond Programming Languages. *IEEE Computer, 16*(8), 57-69.

Schneiderman, B., & Maes, P. (1997). Direct Manipulation vs Interface Agents, Excerpts from the Debates at IUI 97 and CHI 97. *Interactions, 4*(6), 42-61.

Schurr, N., Marecki, j., Scerri, P., Lewis, J. P., & Tambe, M. (2005). The DEFACTO System: Coordinating Human-Agent Teams for the Future of Disaster Response In R. H. Bordini, M. Dastani, J. Dix & A. El Fallah Seghrouchni (Eds.), *Programming Multiagent Systems: Languages, Platforms and Applications* (Vol. 15, pp. 197-215): Springer Press Book Chapter.

Schurr, N., Marecki, J., Tambe, M., Scerri, P., Kasinadhuni, N., & Lewis, J. P. (2005). The Future of Disaster Response: Humans Working with Multiagent Teams using DEFACTO In J. Yen & R. Popp (Eds.), *AI Technologies for Homeland Security: Papers from the 2005 Spring Symposium* Menlo Park, California.: American Association for Artificial Intelligence.

Shanahan, M. (2006). The Frame Problem [Electronic Version]. *The Stanford Encyclopedia of Philosophy*. Retrieved August 2007 from http://plato.stanford.edu/archives/spr2006/entries/frame-problem.

Sheridan, T. B. (1992). *Telerobotics, automation, and human supervisory control*. Cambridge, MA: MIT Press.

Sierhuis, M., Bradshaw, J., Acquisti, A., van Hoof, R., Jeffers, R., & Uszok, A. (2003). *Human-Agent Teamwork Adjustable Autonomy in Practice*. Paper presented at the 7th International Symposium on Artificial Intelligence, Robotis, and Automation in Space.

Simon, H. (1981). *The Sciences of the Artificial* (2e ed.). Cambridge, Massachusetts: The MIT Press.

Simon, H. (1997). Allen Newell: A Biographical Memoir [Electronic Version]. *Biographical Memoirs* 17 from http://www.nap.edu/readingroom/books/biomems/anewell.html.

Sismondo, S. (2004). *An Introduction to Science and Technology Studies*. Oxford, Uk: Blackwell Publishing Ltd.

Skagestad, P. (1993). Thinking With Machines: Intelligence Augmentation, Evolutionary Epistemology, and Semiotic. *The Journal of Social and Evolutionary Systems, 16*(2), 157-180.

Stahl, B. C. (2004). Information, Ethics, and Computers: The Problem of Autonomous Moral Agents *Minds and Machines, 14*, 67-83.

Storms, P. P. A. (2004a). An Agent-Oriented Architecture for Combined Systems. from http://combined.decis.nl/tiki-list_file_gallery.php?galleryId=2

Storms, P. P. A. (2004b). *Combined Systems - A System of Systems Architecture*. Paper presented at the the 1st International Workshop on Information Systems for Crisis Response and Management (ISCRAM), Brussels, Belgium.

Suchman, L. (1987). *Plans and Situated Actions: The Problem of Human/Machine Communication*. New York: Cambridge University Press.

Suchman, L. (1998). Human/machine reconsidered. *Cognitive Studies, 5*(1), 5-13.

Suchman, L. (2003). Figuring Service in Discourses of ICT: the Case of Software Agents. In E. Wynn, E. Whitley, M. Myers & J. DeGross (Eds.), *Global and Organizational Discourse about Information Technology* (pp. 33-43). Boston, Dordrecht: Kluwer Academic Publishers.

Suchman, L. (2008). Feminist STS and the Sciences of the Artificial. In E. J. Hackett, O. Amsterdamska, M. Lynch & J. Wajcman (Eds.), *The Handbook of Science and Technology Studies* (3rd ed., pp. 139-164). Cambridge, Massachusetts: The MIT Press.

Sycara, K. (1998). Multiagent Systems. *AI Magazine, 19*(2), 79-92.

Sycara, K., & Sukthankar, G. (2006). *Literature Review of Teamwork Models*. Pittsburgh, Pennsylvania, USA: Robotics Institute, Carnegie Mellon University.

Traum, D., Rickel, J., Gratch, J., & S., M. (2003). Negotiation over Tasks in Hybrid Human-Agent Teams for Simulation-based Training. In *Proceedings of International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*.

Turing, A. (1950). Computing Machinery and Intelligence. *Mind, 59*, 433-560.

Turkle, S. (1984). *The Second Self: Computers and the Human Spirit*. New York: Simon and Schuster Inc.

Turkle, S. (2005). *The Second Self: Computers and the Human Spirit* (2nd ed.). New York: Simon and Schuster Inc.

van den Hoven, J. (2002). Wadlopen bij Opkomend Tij: Denken over Ethiek en Informatiemaatschappij. In J. de Mul (Ed.), *Filosofie in Cyberspace* (pp. 47-65). Kampen, the Netherlands: Uitgeverij Klement.

van Lente. (1993). *Promising Technology: The Dynamics of Expectations in Technological Development*. Unpublished PhD dissertation, University of Twente, Enschede.

van Lente, H. (2000). Forceful Futures: From Promise to Requirement. In N. Brown, B. Rappert & A. Webster (Eds.), *Contested Futures: A sociology of prospective techno-science* (pp. 43-64). Burlington, VT, : Ashgate Publishing Company.

Verbeek, P. P. (2000). *De Daadkracht der Dingen: over Techniek, Filosofie en Vormgeving*. Amsterdam, The Netherlands: Boom.

Verbeek, P. P. (2006). Materializing Morality. *Science, Technology and Human Values, 31*(3), 361-380.

Verhagen, H. (2003). Autonomy and Reasoning for Natural and Artificial Agents. In M. Nickles, M. Rovatsos & G. Weiss (Eds.), *Agents and Computational Autonomy: Potential, Risks, and Solutions* (pp. 83-94). Berlin; Heidelberg: Springer-Verlag.

Veruggio, G. (2006). *EURON Roboethics Roadmap*. Genoa: Scuola di Robotica.

Vidal, J. (2004, August 5th). The alco-lock is claimed to foil drink-drivers. Then the man from the Guardian had a go ... *The Guardian*.

Vinge, V. (1993). The Technological Singularity. *Whole Earth Review* (Winter issue).

Weiser, M. (1991). The Computer for the 21st Century. *Scientific American 265*(3), 94-102.

Weiser, M. (1993). Some Computer Science Issues in Ubiquitous Computing. *Communications of the ACM, 36* (7), 75-84.

Weiß, G. (Ed.). (1999). *Multiagent Systems: a Modern Approach to Distributed Artificial Intelligence*. Cambridge, Massachusetts: The MIT Press.

West, D., & Travis, L. (1991). The Computational Metaphors and Artificial Intelligence: A Reflexive Examination of a Theoretical Framework. *AI Magazine, 12*(1), 64-79.

Wiener, E. L. (1985). Beyond the Sterile Cockpit. *Human Factors : the Journal of the Human Factors Society, 27*, 75-90.

Wijngaards, N., J. E., Nieuwenhuis, K., & Burghardt, P. (2004). Actor-Agent Communities in Dynamic Environments.

Winograd, T. (2006). Shifting Viewpoints: Artificial Intelligence and Human-Computer Interaction. *Artificial Intelligence, 170*, 1256-1258.

Winograd, T., & Flores, C. F. (1987). *Understanding computers and cognition: a new foundation for design*. Reading, MA [etc.]: Addison-Wesley.

Wood, G. (2002). *Edison's Eve: A Magical History of the Quest for Mechanical Life*. New York: Alfred A. Knopf.

Woods, D. D., Tittle, J., Feil, M., & Roesler, A. (2004). Envisioning Human-Robot Coordination in Future Operations. *IEEE transactions on systems, man and cybernetics, vol. 34 pag. 210-218*(2), pag. 210-218.

Wooldridge, M. (1999). Intelligent Agents. In G. Weiß (Ed.), *Multiagent Systems : a Modern Approach to Distributed Artificial Intelligence* (pp. 27-78). Cambridge, Massachusetts: The MIT Press.

Wooldridge, M. (2002). *An Introduction to Multiagent Systems*. Chichester: Wiley.

Wooldridge, M., & Jennings, N., R. (1995). Intelligent Agents: Theory and Practice. *The Knowledge Engineering Review, 10*(2), 115-152.

Woolgar, S. (1987). Reconstructing Man and Machine: a Note on Sociological Critiques of Cognitivism. In W. Bijker, T. Hughes & T. Pinch (Eds.), *The Social Construction of Technological Systems* (pp. 311-328). Cambridge, Massachusetts: The MIT Press.

Woolgar, S. (1991). Configuring the User: the Case of Usability Trials. In J. Law (Ed.), *A Sociology of Monsters: Essays on Power Technology* (pp. 57-99). London: Routledge.

Wyatt, S. (2008). Technological Determinism is Dead; Long Live Technological Determinism. In E. J. Hackett, O. Amsterdamska, M. Lynch & J. Wajcman (Eds.), *The Handbook of Science and Technology Studies* (pp. 165-181). Cambridge, Massachusetts: The MIT Press.

Zambonelli, F., Jennings, N. R., & Wooldridge, M. (2003). Developing Multiagent Systems: the Gaia Methodology. *ACM Trans on Software Engineering and Methodology 12*(3), 317-370.

Zambonelli, F., & Luck, M. (2004). Agent Hell: A Scenario of Worst Practices. *Computer, 37*(3), 96-98.

Zambonelli, F., & Parunak, H. (2003). Signs of a Revolution in Computer Science and Software Engineering. In *Engineering Societies in the Agents World III* (Vol. 2577, pp. 120-125). Berlin, Heidelberg: Springer-Verlag.

Zambonelli, F., & Parunak, V. (2003). Towards a Paradigm Change in Computer Science and Software Engineering: a Synthesis. *The Knowledge Engineering Review, 18*(4), 329-342.

SUMMARY

Since the 1950's, developments in the field of Artificial Intelligence have brought forth future visions of smart computers and humanoid robots that make life easier and more enjoyable. Such promising visions are offset by dystopian scenarios of worlds in which humans are at the mercy of complex and uncontrollable computer systems. Diverging images like these raise familiar question, like 'will computers ever be able to think?'; 'will they have emotions?' or 'will they be able to love?'. Whether computer will or should be like humans continues to be a key issue in discussions on AI. However, the question is whether this is the right issue to focus on?

This dissertation offers a different perspective on the discussions about whether the development of intelligent technologies will or should bridge the gap between humans and technologies. It explores how the descriptions of these technologies relate to current research and development practices, and in what sense these descriptions can tell us something about future human/technology relationships. The focus in this dissertation is on visions of and research on *intelligent artificial agents*. AI researchers use metaphors to describe the behavior of envisioned computer systems in terms of human features. The metaphor of computer systems as intelligent agents offers a conceptual framework to support at least two ways of thinking about the development of innovative computer technology. First of all, this metaphor provides a conceptual framework for research and development practices. In these practices software and robots are thought of as interactive and social entities that are capable of independently operating in complex and dynamic environments. At the same time, AI researchers and futurologists use this metaphor to construct visions of worlds in which computers move increasingly closer to humans. They present us with images in which intelligent, electronic entities think for and with us, learn, operate independently, and adapt to our needs, habits and preferences. This dissertation shows that metaphors acquire distinct interpretations in these two ways of thinking. A nuanced debate about the possibilities, limitations, and risks of future agent technologies, therefore, should begin with an analysis of the metaphors used to conceptualize these technologies.

In the introductory chapter I outline two departure points for this dissertation. The first concerns the context-dependent nature of human/technology relationships. Too often participants in debates about artificial agents discuss the possibility of creating smart digital entities, without taking into account the context in which these technologies are developed and used. Using the insights from constructivist studies of human/technology relationships, I address the added value of a contextualized perspective on technological development. These studies, which are often grouped under the header Science and Technology Studies (STS), have shown that we have to look beyond inherent properties, in order to understand how humans and technologies relate. The connections between humans and technologies are context-dependent and are shaped by historical, cultural, economic, social and political factors. An analysis of these connections therefore demands a broader perspective of the *sociotechnical* systems in which they are situated.

The second departure point is the role of metaphorical concepts in AI research. Metaphors highlight and hide particular aspects of a concept. In this way, they structure our understanding of concepts as well as our actions, perceptions and expectations. A closer investigation of the metaphorical concepts, used by AI researchers, such as 'intelligence' and 'mind', illustrates that these concepts have different meanings and functions in different contexts. Metaphorically structured descriptions of future AI technologies, isolated from the discourse in which they acquire meaning, therefore offer limited insight into technological developments.

In the final part of the introductory chapter I outline the structure of the dissertation and discuss the concept of 'intelligent artificial agents' and 'agent-based computing' as a field of research. Three metaphorical concepts are recurrent features in descriptions of artificial agents and their relations to humans: 'social interaction', 'adaptive systems', and 'autonomy'. The research in this dissertation analyzes the use of these concepts within three *problem domains*. These are areas in which researchers present the development of intelligent technologies as solution to problems in today's human/technology relationships. The first domain concerns the interaction between humans and technologies on the interface level. The second problem domain concentrates on the human/technology relationship on a cognitive level. The delegation of control, responsibility and accountability provides the central focus in the third problem domain.

In Chapter 2 the emphasis is on those future visions which redefine the human/technology relationship on the level of the interface in terms

of social interaction. In this chapter, the analysis concentrates on the rhetorical aspects of visions of agent technologies. By exploring what these visions represent *and* what they hide, I consider why, despite recurring critiques, the metaphor of social interaction continues to inspire research on intelligent technologies. Social, interactive agents count as optimal solutions for current problems in humans interactions with computers. Today's computer systems are presented as out-dated and not very user-friendly. Making computers more like humans in terms of their communicative skills, interactive behavior and even their appearance would enable a more intuitive and natural interaction. Humans would find it easier and more enjoyable to work with these systems. This chapter shows that such visions offer a narrow view of possible human/technology configurations. Normative choices remain hidden as a result of abstract and often conflicting conceptions of humans, technologies and the relationships between them. Exploring human/technology relationships within concrete contexts of use, exposes the reductionist character of these future visions.

In the third chapter I turn my attention to the instrumental role of metaphorical concepts and their varying interpretations in research on agent technologies. I analyze how researchers in different contexts use and interpret metaphorical concepts to conceptualize human/technology relationships on a cognitive level. In addition, I look at the way in which these concepts influence researchers in the development of these technologies. The emphasis in this chapter is on the use of the metaphor of computer systems as adaptive and self-organizing systems. Adaptive and self-organizing systems should be able to independently adjust to unknown, complex and unpredictable environments. This metaphor is a recurrent element in visions, in which human and technologies figure as two comparable entities, which in the ideal case, relate to each other in a symbiotic way. The notion of symbiotic systems supports a conceptual framework in which humans and technologies can be leveled in terms of cognitive properties.

Metaphorical concepts, such as 'adaptive systems', are not isolated things. They acquire different meanings depending on the context of discourse in which they are used. Therefore, we cannot consider them in isolation from this context, if we want to understand their meaning. To illustrate this I discuss an industrial project, in which researchers are working on the development of an adaptive decision-support system. Within this project metaphorical language fulfills at least two functions. First of all, it has a heuristic function, in the sense that it supports

researchers in the development of computer systems. Secondly, more ambitious visions are presented to contractors and other 'non-experts'. In these visions the development of adaptive agent technologies would, for example, reduce the gap between humans and technologies in terms of their cognitive skills. With the help of these different images, the project is positioned in relation to other technologies and projects. I contrast the two interpretations of adaptive systems with an alternative explanation. In the theory of distributed cognition the differences between humans and technologies play an important role: humans and technologies are complementary elements in cognitive systems. The adaptivity of these systems, this theory tells us, cannot be reduced to the properties of individual components.

The conflicts between different meanings and functions of metaphors and their consequences are explored in Chapter 4. This chapter focuses in particular on the conflicting meanings of the concept of autonomy. The prospect of increasingly complex and autonomous technologies has generated concerns about the delegation and distribution of control, responsibility and accountability between humans and technologies. What happens when things go wrong? Can we still hold humans responsible for accidents caused by independently acting, computer systems? Complex, operating systems mask decision-making processes to such an extent that they are no longer traceable or comprehensible for any one individual. Some agent researchers have argued that these issues can be resolved through the development of autonomous, moral agents. Such agents should be able to reason about the possible moral consequences of their own and human actions. Visions of this kind raise questions, such as 'at what point should we conceive of computers as moral agents?'; 'what would be the consequences?' and 'can we hold computers responsible?'.

A discussion about the possible risks of autonomous agents should begin by asking how we can meaningfully speak about these technologies. Given the metaphorical, context-dependent and instrumental character of concepts used to support visions about future agents, we cannot take claims about their autonomy at face value. The analysis in this chapter focuses on two meanings of autonomy, which are confronted within the discourse on autonomous agents. On the one hand, autonomy is a concept inextricably linked with the notion of a moral and rational person, and rooted in a liberal democratic tradition. On the other hand, autonomy in computer science is a measurable and observable property of the relationship between biological or mechanical

systems and their environments. The tension between these two meanings is indicative of an anthropocentric bias in western democratic societies. In these societies ultimate responsibility is still attributed to humans.

However, the existence of the anthropocentric bias is only part of the story. The tendency to hold humans ultimately morally responsible, rather than machines, does not mean that responsibility is attributed in equal measures to all humans. Not every person is considered to be able or in the position to make moral decisions. Autonomy is malleable and context-dependent concept. Although the anthropocentric bias constrains the space in which technologies perform, technologies in turn set conditions on the range of humans actions, often in ways not anticipated in their design. Technological artifacts persuade, facilitate and enable particular human cognitive processes, actions or attitudes, while constraining, discouraging and inhibiting others. Technological systems, thus, directly influence the autonomy of a person, both in terms of the possibility of a person to act voluntarily, and in terms of the autonomy she is attributed. When we consider these dependencies from a broader sociotechnical perspective, we see that the conceptualization of autonomy structures the organization of the environment in which humans and technologies become connected. It shapes the practices in which humans develop and use technologies. This chapter shows that the concerns about increasingly complex and independently acting computer systems can neither be resolved through abstract analyses of the properties of humans and technologies, nor by developing technologies to be more like humans. Conflicting meanings signal conflicting conceptual and normative frameworks, which should be subject to debate in every context.

In the fifth and final chapter of this book I bring the discussions together and reflect on the implications of my approach both on debates about the possibilities, limitations, and risks of intelligent technologies, as well as on research practices. The research in this dissertation underlines that the promises of envisioned technologies cannot be evaluated on the basis of abstract, decontextualized descriptions. My analysis shows that assumptions, interests and ambitions underlie the choices, interpretations and uses of the agent metaphor. Hence, visions of future agent technologies are best conceived of as narratives about how the relationship between humans and technologies *should be*. They reflect the presuppositions and ideologies of those who construct them. We therefore always have to ask ourselves what descriptions of changing relationships

represent, what they hide, why they are so appealing and where they come from. In addition, an important question is what these visions do in particular contexts. The analysis shows that these visions guide the development in important and different ways. However, it also demonstrates that we can only theorize and empirically study the future connections between humans and technologies to a certain extent. These observations pose the question of how we should structure a debate about future technologies.

The research presented in this book offers instruments for a pragmatic approach to discussions about the possibilities, limitations and risks of intelligent technologies. This approach entails a critical analysis of the metaphorical concepts in research on agent technologies, informed by empirical studies of human/technology relationships. Such an analysis provides a basis for a discussion about the conditions under which technologies *should* or *should not* be considered human-like. It shifts the attention from presumed inevitable trends, towards the choices and conflicts that are associated with particular conceptualizations of human/technology relationships. I discuss how paying attention to these choices and conflicts can contribute both to more reflexive research and development practices, as well as to a broader debate about the social and ethical aspects of agent technologies, and intelligent technologies in general.

Finally, this concluding chapter calls for a broader debate on how we should understand the similarities and differences between technologies and humans within particular contexts. I argue that the boundaries between humans and technologies will sooner be shifted, than dissolved, as a result of a technological development. The key questions then are how and why are these boundaries shifted, and what are the consequences? Who or what is excluded and why? This dissertation shows that the ways in which we understand the gap between humans and technologies in concrete research and development practices is an important topic of discussion, as it is through conceptualizing this gap that we conceptualize and shape ourselves, our technologies and ultimately our society.

NEDERLANDSE SAMENVATTING

Let op de afstand: een kritische analyse van mens/technologie analogieën in discoursen over artificiële agenten

Sinds de jaren vijftig hebben ontwikkelingen op het gebied van de Kunstmatige Intelligentie ('Artificial Intelligence') geleid tot toekomstbeelden waarin slimme computers en mensachtige ('humanoid') robots het leven aangenamer en makkelijker maken. Tegenover deze veelbelovende toekomstbeelden staan dystopische scenario's van werelden waarin mensen zich moeten voegen naar complexe en oncontroleerbare computersystemen. Zulke uiteenlopende visies roepen bekende vragen op: zullen computers ooit kunnen denken? Kunnen ze emoties hebben, of zelfs liefhebben? De vraag of computers als mensen kunnen zijn en wat daar de consequenties van zijn, is een terugkerend element in de discussies over 'Artificial Intelligence' (AI). Is dit echter wel de juiste vraag?

In dit proefschrift benader ik de discussies over het idee dat de ontwikkeling van intelligente technologieën de afstand tussen mens en technologie zal of moet overbruggen vanuit een ander perspectief. Het proefschrift onderzoekt hoe de beschrijvingen van deze technologieën zich verhouden tot de huidige onderzoeks- en ontwikkelpraktijken, en in welk opzicht ze ons iets kunnen vertellen over toekomstige mens/technologie relaties. Toekomstbeelden over en onderzoek naar *intelligente artificiële agenten* ('intelligent artificial agents') staan hierbij centraal. AI onderzoekers beschrijven met behulp van metaforen het gedrag van nog te ontwikkelen computersystemen in termen van menselijke eigenschappen. De metafoor van computersystemen als intelligente agenten biedt een conceptueel kader voor minstens twee manieren van denken over het ontwikkelen van innovatieve computer technologieën. Primair ondersteunt dit beeld het denken en handelen in onderzoeks- en ontwikkelpraktijken. Daarin worden software en robots omschreven als interactieve en sociale entiteiten die in staat zijn om zelfstandig te opereren in dynamische en complexe omgevingen. Daarnaast geven AI onderzoekers en futurologen met behulp van deze metafoor vorm aan toekomstbeelden, waarin computers steeds meer op mensen gaan lijken. Zij spiegelen ons vergezichten voor van toekomstige samenlevingen waarin intelligente, elektronische entiteiten voor en met

ons denken, leren, zelfstandig handelen en zich aanpassen aan de wensen, voorkeuren en gewoontes van mensen. Dit proefschrift laat zien dat metaforen in deze twee manieren van denken op verschillende manieren kunnen worden geïnterpreteerd. Een kwalitatief goed debat over de mogelijkheden, beperkingen en risico's van toekomstige agenttechnologieën vereist daarom, om te beginnen, een analyse van de metaforen die gebruikt worden om deze technologieën te conceptualiseren.

In het inleidende hoofdstuk zet ik de twee vertrekpunten van dit proefschrift uiteen. Het eerste betreft de contextafhankelijkheid van mens/technologie relaties. De discussies over artificiële agenten gaan maar al te vaak over de mogelijkheid om digitale slimme entiteiten te ontwikkelen, zonder dat daarbij gekeken wordt naar de context waarin de betreffende computersystemen worden ontwikkeld en gebruikt. Aan de hand van inzichten die onder meer voortkomen uit constructivistische studies van mens/technologie relaties, bespreek ik de meerwaarde van een gecontextualiseerd perspectief op technologische ontwikkeling. Dit type onderzoek, dat vaak onder de noemer 'Science and Technology Studies' (STS) wordt geschaard, heeft aangetoond dat we verder moeten kijken dan inherente eigenschappen om te begrijpen hoe mensen en technologieën zich verhouden. De verbindingen tussen mens en technologie zijn contextafhankelijk en worden mede beïnvloed door historisch, economisch, sociale, politieke en culturele factoren. Een analyse van deze verbindingen vereist daarom een breder perspectief op de *sociotechnische systemen* waar ze deel van uitmaken.

Het tweede vertrekpunt van dit proefschrift betreft de rol van metaforische concepten in AI onderzoek. Metaforen belichten bepaalde aspecten van een concept en maskeren anderen. Op deze manier structureren ze ons begrip van concepten en daarmee ook ons handelen, onze ervaringen en onze verwachtingen. Een nadere beschouwing van metaforische concepten die AI onderzoekers gebruiken, zoals 'intelligentie' en 'mind', laat zien dat deze concepten in verschillende contexten verschillende betekennissen en functies hebben. Metaforische beschrijvingen van toekomstige AI technologieën, geïsoleerd van het discours waarin ze betekenis krijgen, geven daarom maar een beperkt inzicht in toekomstige technologische ontwikkelingen.

In het laatste deel van het inleidende hoofdstuk ga ik dieper in op het concept van 'intelligent artificial agents' en op 'agent-based computing' als onderzoeksveld, en zet ik de opbouw van het proefschrift uiteen. Onderzoekers van agenttechnologieën gebruiken onder andere drie

concepten om agenten en hun relaties tot mensen te beschrijven: 'sociale interactie', 'adaptieve systemen' en 'autonomie'. Ik analyseer het gebruik van deze concepten binnen drie *probleemdomeinen*. Dit zijn gebieden waarin onderzoekers de ontwikkeling van intelligente agenten presenteren als oplossing voor een hedendaags probleem in de relaties tussen mensen en technologieën. Het eerste domein betreft de interactie tussen mensen en technologie op het interface niveau. Het tweede probleemdomein concentreert zich op de relatie tussen mens en technologie op het cognitieve niveau. In het derde domein staat ten slotte de delegatie en distributie van controle, verantwoordelijkheid en verantwoording centraal. De drie probleemdomeinen bieden een achtergrond voor de drie kernhoofdstukken.

In hoofdstuk 2 ligt de nadruk op toekomstbeelden die de mens/technologie relatie, op het niveau van de interface, herdefiniëren in termen van sociale interactie. De analyse in dit hoofdstuk concentreert zich op de retorisch aspecten van visies over agenttechnologieën. Door te verkennen wat deze visies representeren *en* verbergen, onderzoek ik waarom, ondanks aanhoudende kritiek, de metafoor van sociale interactie onderzoek op het gebied van intelligente technologieën blijft inspireren. Sociale, interactieve agenten gelden als de beste oplossing voor onze omgangsproblemen met computers. Computersystemen van nu worden afgeschilderd als verouderd en niet gebruikersvriendelijk. Door computers meer op mensen te laten lijken in hun communicatieve vaardigheden, hun interactieve gedrag en zelfs hun voorkomen zouden mensen makkelijker met deze systemen om kunnen gaan. Het gebruik zou intuïtiever en natuurlijker worden. Dit hoofdstuk laat zien dat deze toekomstvisie ons een beperkte gunt biedt op mogelijke mens/technologie configuraties. Normatieve keuzes blijven verborgen als gevolg van abstracte en vaak conflicterende concepties van de mens en technologie, en de relatie tussen die twee. Het reductionistische karakter van deze toekomstvisies treedt duidelijk naar voren zodra mens/technologie relaties bekeken worden binnen concrete gebruikscontexten.

In het derde hoofdstuk richt ik me op de instrumentele rol van metaforische concepten en hun verschillende interpretaties in onderzoek naar agenttechnologieën. Ik analyseer hoe onderzoekers in diverse contexten metaforische concepten gebruiken en interpreteren om mens/technologie relaties te conceptualiseren. Tegelijkertijd kijk in naar de manier waarop deze concepten onderzoekers beïnvloeden bij het ontwikkelen van technologieën. De focus ligt hierbij op een analyse in het cognitieve probleemdomein van de metafoor van computersystemen als 'adaptieve

en zelforganiserende systemen'. Adaptieve en zelforganiserende agentsystemen zouden in staat moeten zijn om zich zelfstandig aan te passen aan onbekende, complexe en onvoorspelbare omgevingen. Deze metafoor is een terugkerend element in visies waarin mensen en computertechnologieën figureren als vergelijkbare entiteiten die in het ideale geval in een symbiotische relatie tot elkaar staan. Het idee van een symbiotisch systeem ondersteunt een conceptueel kader waarin de grenzen tussen bepaalde cognitieve eigenschappen van mensen en eigenschappen van computertechnologieën verdwijnen.

Metaforische concepten als adaptieve systemen staan niet op zichzelf. Ze krijgen verschillende betekenissen toegekend, afhankelijk van de context (of het discours) waarin ze worden gebruikt. Om hun betekenis te begrijpen kunnen we ze daarom niet los zien van deze context. Om dit te illustreren bespreek ik een industrieel project waarin wordt gewerkt aan een adaptief, beslissingsondersteunend computersysteem ('decision-support system'). Binnen dit project blijkt metaforisch taalgebruik twee functies te vervullen. Allereerst een heuristische functie, dat wil zeggen: het taalgebruik ondersteunt onderzoekers in het ontwikkelen van computersystemen. Ten tweede, ontstaan al snel weidsere vergezichten in het spreken met opdrachtgevers en andere 'leken'. Zo zou de ontwikkeling van adaptieve agenttechnologieën de afstand tussen mensen en computers in termen van hun cognitieve vaardigheden moeten verkleinen. Met behulp van dit type beelden positioneert men het project te midden van andere technologieën en projecten. Deze twee interpretaties van adaptieve systemen contrasteer ik vervolgens met een alternatieve uitleg. Vanuit de *theorie van gedistribueerde cognitie* spelen de verschillen tussen mens en technologie juist een belangrijke rol: mens en technologie zijn complementaire elementen in cognitieve systemen. Het adaptieve vermogen van deze systemen kan volgens deze theorie niet worden gereduceerd tot de eigenschappen van een enkele component.

De conflicten tussen verschillende betekenissen en functies van metaforen, en de consequenties hiervan staan centraal in hoofdstuk 4. Ik kijk in het bijzonder naar conflicterende betekenissen van het concept autonomie. Het vooruitzicht van steeds autonomere computersystemen heeft tot veel discussies geleid. Hierin spelen vooral de distributie en delegatie van controle, verantwoordelijkheid en verantwoording tussen mensen en intelligente technologieën een grote rol. Complexe, zelfstandig opererende computersystemen maskeren beslissingsprocessen zodanig dat ze niet langer traceerbaar of begrijpelijk zijn voor individuele personen. Wat gebeurt er vervolgens als dingen fout gaan? Kunnen we

mensen nog steeds verantwoordelijk houden voor de ongelukken veroorzaakt door deze complexe computersystemen? Sommige voorstanders van agenttechnologieën hebben suggereren dat de toenemende complexiteit en afnemende controleerbaarheid de ontwikkeling van autonome morele agenten noodzakelijk maakt. Dergelijke agenten moeten zelfstandig kunnen redeneren over de mogelijke morele consequenties van hun eigen acties en over menselijk handelen. Visies als deze roepen vragen op: wanneer kunnen computers gezien worden als morele agenten en wat zijn de consequenties hiervan? Kunnen we computers verantwoordelijk houden?

Een discussie over de mogelijkheden en risico's van autonome agenten begint bij de vraag hoe we betekenisvol over deze technologieën kunnen spreken. Door de metaforische en contextafhankelijke betekennissen van deze concepten, kunnen we de beweringen over autonome agenten niet zonder meer voor waar aannemen. De analyse in dit hoofdstuk richt zich op twee betekennissen van autonomie binnen het vertoog over autonome agenten. Enerzijds is autonomie als ideaal diep geworteld is in westerse democratische tradities. In deze context is het concept onlosmakelijk verbonden met het idee van morele en rationele personen. Anderzijds is autonomie in 'computer science' een meetbare en waarneembare eigenschap van relaties tussen biologische en mechanische systemen en hun omgeving. De spanning tussen deze twee betekenissen in het 'agent discours' duidt op een *antropocentrische bias* in westerse democratische samenlevingen. In deze samenlevingen krijgen mensen als rationele en morele personen nog altijd de eindverantwoordelijkheid toegedicht.

De aanwezigheid van de antropocentrische bias is echter maar een deel van het verhaal. Mensen worden niet te allen tijde gezien als autonome personen. Autonomie is een flexibel en contextafhankelijk concept. De antropocentrische bias beperkt de ruimte waarin computersystemen kunnen handelen, maar tegelijkertijd beïnvloeden computersystemen menslijk handelen, en hoe mensen de wereld ervaren. Technologische systemen hebben daarom een directe invloed op de autonomie van een persoon, zowel op de mogelijkheid van personen om vrij te handelen, als op hoe autonomie wordt toegekend. Als we deze constatering in een breder sociotechnisch perspectief plaatsen, dan zien we dat de conceptualisering van autonomie effect heeft op de organisatie van de omgevingen waarin mens en technologie samenkomen. Het beïnvloedt en vormt de praktijken waarin mensen technologie ontwikkelen en gebruiken. Dit hoofdstuk laat zien dat de zorgen over de steeds com-

plexere en zelfstandig opererende computersystemen niet opgelost kunnen worden door middel van abstracte analyses van de eigenschappen van mensen en technologieën, of door technologieën meer op mensen te laten lijken. Conflicterende betekenissen duiden op conflicterende conceptuele en normatieve kaders, die voor elke context opnieuw onderwerp van discussie moeten zijn.

In het vijfde en laatste hoofdstuk van dit boek breng ik de discussies samen en reflecteer ik op de implicaties van de door mij gekozen aanpak voor het debat over mogelijkheden, beperkingen en risico's van intelligente technologieën en voor onderzoekspraktijken. Het onderzoek in dit proefschrift onderstreept dat de beloftes betreffende toekomstige technologieën niet beoordeeld kunnen worden op basis van abstracte beschrijvingen zijn ontdaan van hun context. Uit mijn analyse blijkt dat veronderstellingen, belangen en ambities ten grondslag liggen aan de keuze, de interpretatie en het gebruik van de agent-metafoor. De vergezichten van toekomstige agenttechnologieën kunnen daarom ook het beste worden gezien als narratieven over hoe de relatie tussen mensen en technologie *zou moeten* zijn. Ze zeggen veel over de overtuigingen en ideologieën van degenen die ze construeren. We moeten ons daarom altijd afvragen wat beschrijvingen van veranderende relaties representeren, wat ze verbergen, waarom ze zo aantrekkelijk zijn en waar ze vandaan komen. Evenzo belangrijk is de vraag wat deze visies 'doen' in bepaalde contexten. Mijn analyse laat zien dat deze visies de ontwikkeling van technologie in belangrijke mate en op verschillende manieren sturen. Uit deze analyse blijkt echter ook dat we maar tot een bepaalde hoogte theoretisch en empirisch kunnen bestuderen hoe mensen en nieuwe technologie uiteindelijk verbonden zullen raken.

Deze constateringen leiden tot de vraag op welke manier we een debat over toekomstige technologieën kunnen vormgeven. Het onderzoek in dit proefschrift biedt instrumenten voor een pragmatische benadering van discussies over de mogelijkheden, beperkingen en risico's van intelligente technologieën. Deze benadering behelst een kritische analyse van de metaforische concepten in onderzoek naar agenttechnologieën, geïnformeerd door empirische studies van de mens/technologie relaties. Een dergelijke analyse biedt een basis voor een discussie over de voorwaarden waaronder technologieën wel of juist als menselijk niet gezien moeten worden. Het richt de aandacht niet op zogenaamde onafwendbare trends, maar op keuzes en conflicten die gepaard gaan met bepaalde conceptualiseringen van mens/technologie relaties. Ik bespreek hoe de aandacht voor deze keuzes en conflicten kan bijdragen

aan zowel reflectieve onderzoeks- en ontwikkelpraktijken, als aan een breder debat over de sociale en ethische aspecten van agenttechnologieën, en intelligente technologieën in het algemeen.

Ten slotte bepleit dit concluderende hoofdstuk een breder debat over de manier waarop we de overeenkomsten *en* verschillen tussen technologie en mensen moeten conceptualiseren, binnen bepaalde contexten. Ik betoog dat grenzen tussen mens en technologie niet verdwijnen als gevolg van technologische ontwikkelingen, maar hoogstens worden verschoven. De vraag is dan hoe en waarom deze grenzen verschoven worden en wat hiervan de consequenties zijn. Wie of wat sluiten ze uit? Dit proefschrift leert dat de conceptualisering van de afstand tussen mensen en technologieën in concrete onderzoeks- en ontwikkelpraktijken een belangrijk onderwerp van discussie moet zijn. Door deze afstand een bepaalde betekenis te geven, vormen we immers ook mens, technologie en uiteindelijk de samenleving.

CURRICULUM VITAE

Merel Noorman (1976, Amsterdam) studied artificial intelligence at the University of Amsterdam in the Netherlands and at the University of Edinburgh in Scotland. Her final thesis ('scriptie') concerned the automated recognition of 3D flexible objects in 2D images. After graduating in 2000 in Amsterdam, she moved back to Edinburgh to do a Master's degree in science and technology studies at the University of Edinburgh. She wrote her Master's thesis on social interactions between humans and intelligent technologies. In 2001 she graduated and was subsequently employed as general manager by the software company VicarVision in Amsterdam, which specializes in computer vision software. She started her PhD project in September 2003 at the Faculty of Arts and Social Sciences of Maastricht University. During her last two years in Maastricht, she represented the PhD students of the faculty in the faculty's research consultation council and in the University's PhD student association Provum. She took part in and completed the PhD program of the national research school WTMC (science, technology and modern culture). In April 2008 she accepted her current position as an advisor for the Dutch Council for Social Development ('Raad voor Maatschappelijke Ontwikkeling') in the Hague.