

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

Citation for published version (APA):

Noorman, M. E. (2009). *Mind the gap : a critique of human/technology analogies in artificial agents discourse*. [Doctoral Thesis, Maastricht University]. Datawyse / Universitaire Pers Maastricht. <https://doi.org/10.26481/dis.20090123mn>

Document status and date:

Published: 01/01/2009

DOI:

[10.26481/dis.20090123mn](https://doi.org/10.26481/dis.20090123mn)

Document Version:

Publisher's PDF, also known as Version of record

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.umlib.nl/taverne-license

Take down policy

If you believe that this document breaches copyright please contact us at:

repository@maastrichtuniversity.nl

providing details and we will investigate your claim.

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 betekenissen 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 probleem-domein 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 probleem-domeinen 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 agententechnologieën. 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 agententechnologieën. Ik analyseer hoe onderzoekers in diverse contexten metaforische concepten gebruiken en interpreteren om mens/technologie relaties te conceptualiseren. Tegelijkertijd kijk ik naar de manier waarop deze concepten onderzoekers beïnvloeden bij het ontwikkelen van technologieën. De focus ligt hierbij op een analyse in het cognitieve probleem-domein 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 betekenissen 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 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 menselijk 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 constatering leidt 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 conceptualisering 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.