Anti-Boxology:
Agent Design in Cultural Context

Phoebe Sengers

August 1998

CMU-CS-98-151

Computer Science Department and
Program in Literary and Cultural Theory
Carnegie Mellon University
Pittsburgh, PA

*Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy.*

Thesis Committee:

Joseph Bates, chair
Camilla Griggers
Jill Fain Lehman
Simon Penny

This work was supported in part by a National Science Foundation Graduate Research Fellowship, and by the Office of Naval Research under grant N00014-92-J-1298.

Keywords: Believable agents, action-selection, behavior-based AI, behavior integration, postmodernism, schizophrenia, cultural studies of science, critical technical practice

Abstract

Artificial Intelligence (AI), the design of technology with attributes that we traditionally associate with living beings, generally follows the broader scientific tradition of focusing on technical problems and their solutions within a relatively constrained framework. The cultural studies of science, on the other hand, insists that scientific work should be understood as it springs from and influences other cultural phenomena, including the background of metaphors and assumptions that influence the way scientists do their work. In this thesis, I explore the possibilities for AI and the cultural studies of science to engage in a mutually beneficial alliance, by studying AI as a culturally situated activity and by using results of that study to generate novel technology.

Specifically, I focus on the design of *autonomous agents*, programs which are intended to represent a complete person, animal, or character. In the alternative AI tradition, these agents are created from a set of independent building blocks termed *behaviors*. A major open question is how these behaviors can be synthesized to create an agent with overall coherent behavior. I trace the problems in behavior integration to a strategy called *atomization* that AI shares with industrialization and psychiatric institutionalization. Atomization is the process of breaking agents into modular chunks with limited interaction and represents a catch-22 for AI; while this strategy is essential for building understandable code, it is fatal for creating agents that have the overall coherence we have come to associate with living beings.

I tackle this problem of integration by redefining the notion of agent. Instead of seeing agents as autonomous creatures with little reference to their sociocultural context, I suggest that agents can be thought of in the style of cultural studies as a form of communication between the agent's designer and the audience which will try to comprehend the agent's activity. With this metaphor as a basis, it becomes clear that we need to integrate, not the agent's internally defined code, but the way in which the agent presents itself to the user. Narrative psychology suggests that agents will be maximally comprehensible as intentional beings if they are structured to provide cues for narrative. I therefore build an agent architecture, the *Expressivator*, which provides support for narratively comprehensible agents, most notably by using behavioral transitions to link atomic behaviors into narrative sequences.

Acknowledgements

This thesis is dedicated with gratitude to my family, who gave me support, encouragement, and many entertaining hours of sharp-tongued wit during the years it took to do this work. It is especially dedicated to Nishka, my beloved companion throughout my education, who died, sadly, only a few short weeks before it was over.

I have been fortunate at Carnegie Mellon to be surrounded by many lively, intelligent, and stimulating friends and colleagues. First and foremost, I want to thank my advisors, Joe Bates and Camilla Griggers, to whom I am in great debt.

I was lucky to fall into Joe's lap when I came to CMU. He is one of the smartest people I know, but for me his unique gift in a university full of smart people is that he understands and values the connection between work and life. He is a visionary, and as such knows that work should not only go well but *feel* right, and the kinds of things a person can do to find that right work and go about it, happily. For years, Joe has helped me in innumerable ways to follow my dream of synthesizing AI and cultural studies, even when at times the work that involved seemed to him to be incomprehensible or even bizarre. Things would have gone very differently without him, and I am happy, at the end, to look back and see the firm stamp of his thinking in my work.

Camilla has been a fabulous resource, challenging me to strengthen the theoretical aspects of the dissertation while (perhaps more importantly) helping me learn how to do this work without going insane. two of her greatest gifts to me have been guidance in figuring out what *I* want to do and support in figuring out how to do it. She is also the only advisor I know who, when confronted with a stressed-out student, would give them a foot massage. Camilla, thanks for the support of mind, body, and soul!

I am also grateful to my other committee members, Jill Fain Lehman and Simon Penny, who provided me with valuable comments and advice. Jill's views on AI are both sharp as a tack and very different from mine, and as a consequence her feedback — not to mention her patience with a research area that at times must have seemed from another planet — was enormously helpful. This was Simon's first Ph.D. committee, a fact that was made clear by his sheer enthusiasm; he has not yet learned that a committee member's proper function is to try to avoid interaction with his or her student until the last possible moment. Many thanks to you, Simon, for the frequent meetings over cups of home-grown tea, the conversation that ranged from technical minutiae to theoretical extravaganzas to comments on the contemporary political scene, for the demonstration of the depths to which bad taste in clothing can sink, and for the truly hideous green, orange, and purple polyester dress that accompanied your last set of comments.

Great thanks go to all those who bravely served on my interdisciplinary Ph.D. program committee at a time when it was not at all clear where this research direction would go and whether it was capable of bearing fruit: Merrick Furst, Ronald Judy, Paul Hopper, Nancy Spivey, Kris Straub, and Dave Touretzky. I would particularly like to thank Ronald Judy for having done most of the legwork to get the program approved by the English department. His attention to the details of how the program should be structured and approved saved me much heartache and headache further down the road.

For help with this thesis, I would like to thank Paul Vanouse and Colin Piepgras, for advice on the design of the Industrial Graveyard; James Lester, for help with the design of the user study, which unfortunately never came into being; Bruce Blumberg, for general encouragement and especially for fed exing me a copy of the ALIVE video before my thesis proposal; Kerstin Dautenhahn, for special support and encouragement; and Bruce Blumberg, Rod Brooks, James Lester, and Luc Steels for permission to use their images in this thesis.

One of the greatest pleasures of being a graduate student in LCT is being immersed in an environment of smart, kooky, funky intellectuals whose greatest pleasure after a couple pints of Iron City is the thorough discussion of such issues as the structure of postmodern consciousness as reflected in contemporary advertising for feminine hygiene products. Thanks to my fellow grad students for the moral support, the stimulating and sometimes bizarre conversation, and the happening parties, some of which I will never forget and some of which I still can't remember. Special thanks are also due to honorary grad student Sharon Ghamari-Tabrizi, an inspiring teacher, ferocious intellect, and all-around target of my admiration.

In CS, I was fortunate to work in the context of the Oz Project, a tolerant and encouraging environment for work on the cutting edge of respectability. I am particularly grateful to Bryan Loyall, for many conversations about agent architectures and life; to Scott Neal Reilly, at once the most normal and the most twisted Oz-ite and therefore an infinite source of amusement; to Peter Weyhrauch, for an entire day in Toronto where he sang every sentence, opera-style; and to Michael Mateas, for letting me get to know him before he is really, really famous (I look forward to selling the tell-all memoir). I am also grateful to Catherine Copetas, for her competence in dealing with every problem, no matter how strange, I ever came to her with; and to Sharon Burks, for knowing all the rules and how to bend them in my favor.

Thanks to those who started me on the road to this thesis: to Hank Dardy, who trustingly gave me a research job in Computer Science when I barely knew how to program; to Simon Kasif, who herded me through a rigorous, theoretical undergraduate program, only to discover to his despair that I went off the deep end in graduate school; to Fritz Gutbrodt, who helped me put together the first inklings of what a synthesis of the humanities and technology could be like.

A graduate student's quality of life is extraordinarily affected by the officemates with whom s/he spends a large portion of his or her waking hours. I have been blessed to share my office with a group of inimitable characters and excellent friends: James Landay, the trash talker par excellence; Belinda Thom, the Queen B; and Dayne Freitag, international man of mystery. Y'all are the greatest!

I have been surrounded by many loving friends over my years at CMU. I am grateful to all of them for making these some of the best years of my life. I would especially like to thank the following people: Mary Tomayko, for many late-night phone calls and for being utterly unlike me and still my best friend; Kristin Nelson, for her pig-headedness, which

I adore, and for her sharp wit and willingness to tell it like it is; Jean Camp and Shaun McDermott, for being my second home and family; Nick Thompson, for being himself, an endless source of fascination and object of affection; Laura Ruetsche, for simultaneously being both the smartest and the least pretentious person I know, for her dry wit, and for her ability to blurt out the most unexpected comments at any tme; Stephanie Byram, for her unique and wise outlook on life and for glowing and exotic trip reports; Faye Miller, for her blistering forehand and for many companionable hours of thesis avoidance in her office in Wean Hall; Thorsten Joachims, for his steady nerves, for his unflagging enthusiasm for unexpected trips, unusual activities, and anything vaguely edible, and for his normally well-hidden bizarre side; Doug Davis, for a steady infusion of cheese and overstimulated intellect in my life; and Robbie Warner, for not being my kid and instead being my friend.

Contents

If you embrace the virtual life, don't do it mindlessly;
read what the best critics have to say.
— Howard Rheingold

Chapter 1

Introduction: Agents in Culture

Artificial Intelligence (AI) has come a long way. Particularly in the last ten years, the subfield known as 'agents' — artificial creatures that 'live' in physical or virtual environments, capable of engaging in complex action without human control — has exploded [Johnson, 1997] [Sycara and Wooldridge, 1998][1]. We can now build agents that can do a lot for us: they search for information on the Web [Shakes *et al.*, 1997], trade stocks [Analytix Inc., 1996], play grandmaster-level chess [Hsu *et al.*, 1990], patrol nuclear reactors [Baker and Matlack, 1998], remove asbestos [Schempf, 1995], and so on. Agents have come to be powerful tools.

But one of the oldest dreams of AI is the 'robot friend' [Bledsoe, 1986], an artificial being that is not just a tool but has its own life. Such a creature we want to talk to, not just to find out the latest stock quotes or the answer to our database queries, but because we are interested in its hopes and feelings. Yes, we can build smart, competent, useful creatures, but we have not built very many that seem complex, robust, and alive in the way that biological creatures do. Who wants to be buddies with a spreadsheet program, no matter how anthropomorphized? Somehow, in our drive for faster, smarter, more reliable, more useful, more profitable artificial agents, it seems like we may have lost something equally important: the dream of a creature which is, on its own terms, alive.

At the same time, as the notion of 'agent' has started to take on pop culture cachet, outside academics have begun to turn a not-always-welcome critical eye on the practices of AI. To humanists interested in how AI fits into broader culture, both the goals and the methodologies of AI seem suspect. With AI funding coming largely from the military and big business, critics may wonder if AI is just about building autonomous fighter pilots, more complex voicemail systems, and robots to replace human workers on assembly lines. The notion of the hyperrational, disembodied agent which still drives much AI research strikes many critics as hopelessly antiquated and even dangerous. AI research, these

[1]The format for citations in this thesis is as follows: [Smith, 1998] cites a particular work; ([Smith, 1998], 14) cites a particular page in a particular work; and (14) cites a particular page in the most recently mentioned work.

critics say, is about reproducing in silicon ideas of humanity that are hopelessly limited, leaving out much of what we value in ourselves. AI, in this view, is bad science and bad news.

These critiques, while not always equally easy for AI researchers to hear, could potentially help AI researchers develop better technical practices. They often focus on what has been left out of AI, helping us understand at a deep level why we have not yet achieved the AI dream of artificial creatures that are meaningfully alive, giving us a glimpse of the steps we could take towards fulfilling that dream, and advising us on integrating the practice of AI responsibly with the rest of life. Unfortunately, however, while being eloquent additions to such fields as anthropology, philosophy, or cultural studies, the critiques have often been unintelligible to AI researchers themselves. Lacking the context and background of humanist critics, researchers often see humanist concerns as silly or beside the point when compared to their own deep experiential knowledge of technology. Similarly, humanist critics have generally lacked the background (and, often, the motivation) to phrase their criticisms in ways that speak to the day-to-day technical practices of AI researchers. The result is the ghettoization of both AI and cultural critique: technical practices continue on their own course without the benefit of insight humanists could afford, and humanists' concerns about AI have little effect on how AI is actually done.

The premise of this thesis is that things can be different. Rather than being inherently antagonistic, AI and humanistic studies of AI in culture can benefit greatly from each other's strengths. Specifically, by studying AI not only as technology but also as a cultural phenomenon, we can find out how our notions of agents spring from and fit into a broader cultural context. Reciprocally, if the technology we are currently building is rooted in culturally-based ways of thinking, then by introducing new ways of thinking we can build new and possibly better kinds of technology.

This insight — that cultural studies of AI can uncover groundwork for new technology — forms the basis of this thesis. In particular, I look at methods for constructing artificial creatures that combine many forms of complex behavior. I analyze the technical state of the art with a cultural studies of science perspective to discover the limitations that AI has unknowingly placed upon itself through its current methodological presuppositions. I use this understanding to develop a new methodological foundation for an AI that can combine both humanistic and engineering perspectives. Finally, I leverage these insights in the development of agent technology, in order to generate agents that can integrate many behaviors while maintaining intentional coherence in their observable activity; or, colloquially speaking, appear more *alive*.[2]

But let's start at the beginning, with you and what you bring to this work. You may be an AI researcher, curious about the humanities or only interested in technology. You may be a cognitive scientist, a cultural critic, an anthropologist, a historian, an artist, all of these, some of these, none of these. You may be dying to know how to construct functional agents out of many behaviors; or you may be mildly curious about how AI has imported and modified methodologies from the industrial revolution. You may be a true believer in this interdisciplinary direction or you may

[2] What this means concretely will be made clear in Chapter 6.

be a die-hard skeptic.

In all these cases, this thesis has something to say to you, but in none of them can it do so without your help. This is a thesis which lives in the gap between two disciplines, AI and the cultural studies of science, which share almost nothing in their presuppositions, methodologies or values. As such, it is likely to please no one. If it is seen as a monolithic argument, to be accepted or rejected in its entirety, it will almost inevitably fail.

Instead, I would suggest that you try thinking of it as a toolbox of interconnected ways of thinking, each of which will be more or less useful to you depending on what you do now and what you want to use it for. If you can use the technology but find the philosophy on which it is based implausible, more power to you. If you appreciate the analysis of construction of knowledge about agents, but find the technical application deeply wrong-headed, that's OK too. But you will probably get the most out of this thesis if you find a way to make some sense of even the alien parts of this thesis.

In the rest of this introduction, I will try provide the background knowledge that you will likely need to feel at home in the rest of the thesis. I will introduce the fields of autonomous agents and cultural studies of science. I will give an overview of how agent research and broader culture are intimately intertwined. Then, I will explain how agent research and cultural studies have been profitably combined in the past, and how the approach for synthesizing them provided in this thesis grows out of these past traditions. This will set us up to delve into technical work in Chapter 2.

Introduction to Autonomous Agents

One of the dreams of AI is the construction of independent artificial beings. Rather than slavishly following our orders, or filling some tiny niche of activity that requires some aspect of intelligence (for example, playing chess), these artificial creatures would lead their own existences, have their own thoughts, hopes, and feelings, and generally be independent beings just as other people or animals are. In the 1950's and early 1960's, this dream for AI was embodied in cybernetics. For example, Walter built small robots with rudimentary "agenty" behaviors [Walter, 1963]. He called his robots 'turtles;' they would roam around their environment, seeking light, finding food, and avoiding running into things. Later models could do some rudimentary associative learning.

But as cybernetics fell out of fashion, AI research began to focus more on the cognitive abilities an artificial agent might need to have higher-level intelligence, and less on building small, complete (if not so smart) robots. At least partially because the task of reproducing a complete creature has been so daunting, AI spent quite a few years focused on building individual intelligent capabilities, such as machine learning, speech recognition, story generation, and computer vision. The hope was that, once these capabilities were generated, they could be combined into a complete agent; the actual construction of these agents was often indefinitely deferred.

More recently, however, the field of autonomous agents has been enjoying a renaissance. The area of autonomous agents focuses on the development of programs that more closely approach representations of a complete person or creature. These agents are programs which engage in complex activity without the intervention of another program or person. Agents may be, for example, scientific simulations of living creatures [Blumberg, 1994], characters in an interactive story [Bates, 1994], robots who can independently explore their environment [Simmons *et al.*, 1997], or virtual 'tour guides' that accompany users on their travels on the World Wide Web [Joachims *et al.*, 1997].

While these applications vary wildly, they share the idea that the program that underlies them is like a living creature in some important ways. Often these ways include being able to perceive and act on their (perhaps virtual) environment; being autonomous means they can make decisions about what to do based on what is happening around them and without necessarily consulting a human for help. Agents are also often imputed with rationality, which is defined as setting goals for themselves and achieving them reasonably consistently in a complex and perhaps hostile environment.

Agent as Metaphor

The definition of what exactly is and is not an agent has at times been the source of hefty controversy in the field. Mostly these controversies revolve around the fact that any strictly formal definition of agenthood tends to leave out such well-beloved agents as cats or insects, or include such items as toasters or thermometers that a lay person would be hard-pressed to call an agent. With some of the looser definitions of agents, for which the word 'agent' just seems to be a trendy word for 'program,' skeptics can be forgiven for wondering why we are using this term at all.

In this thesis, I will take agenthood broadly to be a sometimes-useful way to frame inquiry into the technology we create. Specifically, agenthood is a metaphor we apply to computational entities we build when we wish to think of them in ways similar to the ways we understand living creatures. Calling a program an agent means the program's designer or the people who use it find it helpful or important (or, for that matter, attractive to funders to think of the program as an independent and semi-intelligent coherent being. For example, when we think of our programs as agents we focus our design attention on 'agenty' attributes we would like the program to have: the program may be self-contained; it may be situated in a specific, local environment; it may engage in 'social' interactions with other programs or people.[3] When a program is presented to its user as an agent, we are encouraging the user to think of it not as a complex human-created mechanism but as a user-friendly, intelligent creature. If 'actually' some kind of tool, the creature is portrayed as fulfilling its tool-y functions by being willing to do the user's bidding [Lanier, 1996] [Wise, 1996]. Using the metaphor 'agent' for these applications lets us apply ideas about what living agents such as dogs, beetles, or bus drivers are like to the design and use of artificially-created programs.

[3] I am indebted to Filippo Menczer for this observation.

Agenthood in Classical and Alternative AI

But not all AI researchers agree on which conceptions of living agents are appropriate or useful for artificial agents. The past 10 years in particular have seen an at times spectacular debate between different strains of thought about the proper model of agent to use for AI research (see e.g. [Cognitive Science, 1993]). Rodney Brooks [Brooks, 1990], for instance, divides the field into symbolically-grounded vs. physically-grounded agents. Agents based on symbols work by manipulating representations of their environment; physically-based agents work by manipulating and reacting to the environment itself. Philip Agre and David Chapman [Agre and Chapman, 1990] distinguish agents using 'plans-as-programs' from agents using 'plans-as-communication;' they divide programs into ones that engage in abstract, hierarchical planning of activity before engaging in it (often including formal proofs that the plan will fulfill the goal the agent is given) versus ones that are designed to take advantage of an action loop with respect to their environment and may only refer to plans as ways to structure common activities. Another common distinction is between situated and cognitive agents; situated agents are thought of as embedded within an environment, and hence highly influenced by their situation and physical make-up, whereas cognitive agents engage in most of their activity at an abstract level and without reference to their concrete situation.

These divisions are not independent; rather, they tend to repeat similar categories with different names. Specifically, these rubrics tend to organize themselves into two conceptual clusters: a main stream often termed *classical* AI (also known as Good Old-Fashioned AI, cognitivistic AI, symbolic cognition, top-down AI, knowledge-based AI, etc.) and an oppositional stream we can term *alternative* AI (also known as new AI, nouvelle AI, ALife, behavior-based AI, reactive planning, situated action, bottom-up AI, etc.).[4] Not every AI system neatly falls into one or the other category — in fact, few can be said to be pure, unadulterated representatives of one or other. But each stream represents a general trend of thinking about agents that a significant number of systems share.

For AI researchers, the term classical AI refers to a class of representational, disembodied, cognitive agents, based on a model that proposes, for example, that agents are or should be fully rational and that physical bodies are not fundamentally pertinent to intelligence. The more extreme instances of this type of agent had their heyday in the 60's and 70's, under a heady aura of enthusiasm that the paradigms of logic and problem-solving might quickly lead to true AI. One of the earliest examples of this branch of AI is Allen Newell and Herbert Simon's GPS, the somewhat optimistically titled "general problem solver." This program proceeds logically and systematically from the statement of a mathematical-style puzzle to its solution [Newell and Simon, 1972]. Arthur Samuel's checker player, one of the first programs that learns, attempts to imitate intelligent game-playing by learning a polynomial function to map aspects of the current board state to the best possible next move [Samuel, 1995]. Terry Winograd's SHRDLU maintains a simple representation of blocks lying on a table, and uses a relatively

[4] For similar analyses, see e.g. [Steels, 1994] [Varela *et al.*, 1991] [Brooks, 1990] [Norman, 1993].

straightforward algorithm to accept simple natural language commands to move the virtual blocks [Winograd, 1972]. While the creators of these programs often had more subtle understandings of the nature of intelligence, the programs themselves reflect a hope that simple, logical rules might underlie all intelligent behavior, and that if we could discover those rules we might soon achieve the goal of having intelligent machinery.

But the classical model, while allowing programs to succeed in many artificial domains which humans find difficult, such as chess, unexpectedly failed to produce many behaviors humans find easy, such as vision, navigation, and routine behavior. The recognition of these failures has led to a number of responses in the 80's and 90's. Some researchers — most notably Winograd, who wrote an influential book with Fernando Flores on the subject [Winograd and Flores, 1986] — have decided that the intellectual heritage of AI is so bankrupt they have no choice but to leave the field. By far the majority of AI researchers have remained in a tradition that continues to inherit its major research framework from classical AI, while expanding its focus to try to incorporate traditionally neglected problems (we might call this 'neo-classical AI'). A smaller but noisy group has split from classical AI, claiming that the idea of agents that classical AI tries to promote is fundamentally wrong-headed.

These researchers, who we will here call alternative AI, generally believe that the vision of disembodied, problem-solving minds that explicitly or implicitly underlies classical AI research is misguided. Alternative AI focuses instead on a vision of agents as most fundamentally nonrepresentational, reactive, and situated. Alternative AI, as a rubric, states that agents are situated within an environment, that their self-knowledge is severely limited, and that their bodies are an important part of their cognition.

Technology as Theory of Subjectivity

The dialogue and debate between these two types of agents is not only about a methodology of agent-building. An underlying source of conflict is about which aspects of being human are most essential to reproduce. Classicists do not deny that humans are embodied, but the classical technological tradition tends to work on the presupposition that problem-solving rationality is one of the most fundamental defining characteristic of intelligence, and that other aspects of intelligence are subsidiary to this one. Likewise, alternativists do not deny that humans can solve problems and think logically, but the technology they build is based on the assumption that intelligence is inherent in the body of an agent and its interactions with the world; in this view, human life includes problem-solving, but is not a problem to be solved.

It is in these aspects of AI technology — ones that are influenced by and in turn influence the more philosophical perspectives of AI researchers — that we can uncover, not just the technology of agents, but also theories of agenthood. Two levels of thought are intertwined in both these approaches to AI: (1) the level of day-to-day technical experience, what works and what doesn't work, which architectures can be built and which can't; and (2) the level of background philosophy — both held from the start and slowly and mostly unconsciously imbibed within the developing technical traditions — which underlies the way in which the whole complex and undefined conundrum of recreating life in the computer is understood. Running through and along with the technical arguments are more philosophical arguments about what human life is or should be like, how we can come to understand it, what it means to be meaningfully alive.

Technical researchers may feel uncomfortable making this connection between technology and fuzzier ideas about what it is to be human. But this is not revolutionary; the connection between living and artificial agent is ingrained in AI, allowing the connection between AI and psychology that forms cognitive science. For example, when Newell describes his concept of the 'knowledge level' — a way of comprehending beings as agents, rather than as physical organisms or computers — he means for this way of thinking to describe *both* artificial *and* living agents [Newell, 1981]. Both these kinds of agents are described using the same kind of structure: "an agent is composed of a set of actions, a set of goals, and a body... [T]he agent processes its knowledge to determine the actions to take. Finally, the behavior law is the *principle of rationality*. Actions are selected to attain the agent's goals" (13). For Newell, at the knowledge level, an agent is defined to consist of actions, goals, and body; for an entity to be considered an agent, its actions must be oriented to achieving goals. These attributes of agents are considered to hold whether we are talking about computers or people. The knowledge-level theory implies that both kinds of agents are fundamentally structured so that their behavior consists of rational attempts to achieve plausible goals.[5]

But even researchers who do not claim to be doing cognitively plausible work draw their inspiration in part from theories of living agents. This is demonstrated, for example, by the very title of Brooks' position paper opposing classical AI, *Elephants Don't Play Chess*. While Brooks does not claim to be building structures isomorphic to ones inside the mind, he does think that considerations of what 'real' agents do in the world are part of the consideration that should go into the design of an alternative agent. Here, he claims that rational, symbolic, problem-solving behavior is inessential to an agent's existence in the world, which is rather dominated by the need for perception and reactivity.

Cultural theorists use the term 'subjectivity' to refer to theories or models of consciousness. A theory of subjectivity suggests what existence is like, how we come to experience ourselves and the world around us, what it feels like or means to be a person. From the previous discussion, it seems clear that AI includes not only conflicting theories of technology but also, implicitly, conflicting theories of subjectivity. Specifically, classical AI technology is based on a model of

The term 'subjectivity' is related to the perhaps more familiar term 'subjective' in that they both refer to personal experience. 'Subjective' knowledge is something that is known to you as an individual, whereas 'objective' knowledge can be thought of as something that would hold true for anyone, and is therefore not related to or dependent on your life experience.

[5] By 'rationality,' AI researchers often mean 'bounded rationality,' i.e. that the rationality of an agent's behavior is limited to its (presumably limited) knowledge. What I mean to get at here is *not* that the knowledge-level theory implies that computers and people are hyperrational (and perhaps by extension hyperintelligent). Rather, I argue that setting up rationality as one of the *fundamental* characteristics by which agentiness can be defined means that agents which are behaving *irrationally* (as humans often do) are *flawed* in their agenthood. Of course, this irrationality can be, and in AI often is, redefined as rationality with flawed knowledge or in the pursuit of perverted goals, such as that a person is, for example, rationally trying to harm him- or herself — a redefinition that handily circumvents having to deal with the still unanswered (and perhaps unanswerable) question of whether rationality should be considered a fundamental, defining property of the experience of being in the world.

consciousness as essentially representational, rational, and disembodied. Alternative AI technology presupposes that it is essentially reactive, situated, and embodied.

These two categories can be clearly seen within AI research. Within that research community, they are generally seen as coming about from certain tensions in technical practice itself. Interestingly, they correspond closely to two categories cultural theorists regularly employ to talk about historical notions of what it means to be a person: *rational* and *schizophrenic* subjectivity.[6]

Rational subjectivity refers to a common way of conceiving humanity in the West since the Enlightenment. It is historically anchored in the work of René Descartes, the Enlightenment philosopher who derives proof of his existence from the fact that he thinks. Rational subjectivity is based on this Cartesian focus on logical thought: the mind is seen as separated from the body, it is or should be fundamentally rational, and cognition divorced from emotion is the important part of experience. This model has overarching similarities with, for instance, Allen Newell's theory of Soar, which describes an architecture for agents that grow in knowledge through inner rational argumentation [Newell, 1990]. Most models built under Soar are focused on how this argumentation should take place, leaving out issues of perception and emotion (though there are certainly exceptions; see e.g. [Pearson *et al.*, 1993]).

The development of the notion of schizophrenic subjectivity is based on perceived inadequacies in the rational model, and is influenced by but by no means identical to the psychiatric notion of schizophrenia (we will discuss this relationship in more detail in Chapter 2). While rational subjectivity presupposes that people are fundamentally or optimally independent rational agents with only tenuous links to their physicality, schizophrenic subjectivity sees people as fundamentally social, emotional, and bodily. It considers people to be immersed in and to some extent defined by their situation, the mind and the body to be inescapably interlinked, and the experience of being a person to consist of a number of conflicting drives that work with and against each other to generate behavior. In AI, this form of subjectivity is reflected in Brooks's subsumption architecture, in which an agent's behavior emerges from the conflicting demands of a number of loosely coupled internal systems, each of which attempts to control certain aspects of the agent's body based almost entirely on external perception rather than on internal cogitation [Brooks, 1986a].

Each class of agent architectures closely parallels a model of subjectivity. Just as alternative AI has arisen in an attempt to address flaws in classical AI, the schizophrenic model of subjectivity has arisen in response to perceived flaws in the rational model's ability to address the structure of contemporary experience. Each style of agent architecture shows a striking similarity to a historical model of subjectivity that cultural theorists have identified.

This close relationship between a technical debate in a subfield of computer science and philosophical trends in Western culture as a whole may come as a surprise. But a moment of reflection reveals where the

[6]This idea is a more common observation among cultural theorists who study AI. See, for example, [Barton, 1995] and [de Mul, 1997].

connection lies. AI researchers are also human beings, and as such inhabit and are informed by the broader society that cultural theorists study. From this point of view, AI is simply one manifestation of culture as a whole. Its technical problems are one specific arena where the implications of ideas that are rooted in background culture are worked out.

But if AI is fundamentally embedded in and working through culture, then cultural studies and AI may have a lot to say to each other. Specifically, the practitioners of cultural studies — who we will here refer to as cultural critics or cultural theorists — have spent a lot of time thinking about and debating subjectivity. AI researchers have spent a lot of time thinking about and debating architectures for autonomous agents. Once these two are linked, each body of work can be used to inform the other. If agents use a particular theory of subjectivity, then we can use ideas about this theory to inform our work on agents. And if agents are a manifestation of a theory of subjectivity, then studying these agents can give us a better idea of what that theory means. In order to make this idea concrete, we will now look at cultural studies and its relationships to science in more detail.

Cultural Studies Meets Science

Cultural studies — and its related philosophy, cultural theory — is a hybrid collection of literary scholars, anthropologists, philosophers, sociologists, historians, and other sympathetic humanists. While cultural theorists are heterogeneous in both method and philosophy, they generally aim to understand human experience as it is formed and expressed through a variety of cultural forms. A common interest of cultural theorists is understanding how the structure of society both constrains and enables human understanding of ourselves and each other.

One way of understanding the mindset of cultural studies is by looking at how it has grown out of literary studies. Literary studies originally were confined to high literature, i.e. stories by writers such as Shakespeare who are acknowledged as great. Over time, literary scholars began to apply the methods of literary studies to 'low' literature as well, for example dime store novels, as well as works by authors outside the main traditions of Western culture. Soon, these scholars noticed that the same techniques also worked for film, leading to film studies. Gradually, the field expanded to cover all forms of cultural production, including television, advertising, law, politics, religion, and science.

The cultural studies of science — also termed cultural critique of science or science studies — aim to understand science as it relates to the culture of which it is a part. It broadly functions as a kind of 'science criticism' analogous to literary criticism [Harding, 1994]: one of its major goals is to understand and improve the quality and relevance of scientific work by thinking about how it stems from and affects the rest of culture. Science, too, has an ideal of improving itself by continuously subjecting itself to rigorous self-criticism [Rouse, 1993]. But like philosophy of science, science studies aims to understand and improve not only particular technical work, but also the very mechanisms by which science works and through which it produces knowledge. Science studies goes beyond both science and philosophy of science by relating scientific methodology and assumptions to other cultural practices which many scientists and philosophers see as external or irrelevant to the production of science.

Science Wars

Science studies examines culturally-based metaphors that inform scientific work, and thereby often uncovers deeply-held but unstated assumptions that underly it. Scientists are also generally interested in understanding the forces, both conscious and unconscious, that can shape their results. If there are ways in which they can better understand the phenomena they study or build the technology they want to create, they are all ears. In this respect, the insights of science studies can contribute great value to science's self-understanding [Keller, 1985].

At the same time, many practitioners of science studies are deeply interested in science as it is actually practiced on a day-to-day level. This means scientists, with their in-depth personal experience of what it means to do scientific work, are privy to perspectives that can enrich the work of their science studies counterparts. Science studies simply is not possible without science, and an important component of it is an accurate reflection of the experiences of scientists themselves.

With all the advantages that cooperation could bring, you might think that science and science studies would be enthusiastic partners on the road to a shared intellectual enterprise. Alas, that is far from the case! Unfortunately, productive exchanges between cultural critics and scientists interested in the roots of their work are hampered by the disciplinary divide between them [Snow, 1969]. This divide blocks cultural critics from access to a complete understanding of the process and experience of doing science, which can degrade the quality of their analyses and may lead them to misinterpret scientific practices. At the same time, scientists have difficulty understanding the context and mindset of critiques of their work, making them unlikely to consider such critiques seriously or realize their value for their work, potentially even leading them to dismiss all humanistic critiques of science as fundamentally misguided [Gross and Levitt, 1994].

This feedback loop of mutual misunderstanding has grown into a new tradition of mutual kvetching. Cultural critics may complain that scientists unconsciously reproduce their own values in their work and then proclaim them as eternal truth. They may feel that scientists are not open to criticism because they want to protect their high (relative to the humanities') status in society. Simultaneously, scientists sometimes complain that cultural critics are absolute nihilists who do not believe in reality and equate science with superstition.[7] They fear that cultural critics undermine any right that science has as a source of knowledge production to higher status than, say, advertising. Finally, both sides complain incessantly — and correctly —- of being cited, and then judged, out of context.

The unfortunate result of this situation is a growing polarization of the two sides. In the so-called "Science Wars" [Social Text, 1996], pockets

[7]This is exacerbated by the fact that the notion of 'reality' used by many scientists in their criticism of science studies does not bear much relation to the long and deep tradition of the usage of that term in cultural studies of science.

of fascinating interdisciplinary exchanges and intellectually illuminating debate are sadly overwhelmed by an overall lack of mutual understanding and accompanying decline of goodwill. While most participants on both sides of the divide are fundamentally reasonable, communication between them is impaired when both sides feel misunderstood and under attack. This siege mentality not only undermines the possibility for productive cooperation; with unfortunate frequency, it goes as far as cross-fired accusations of intellectual bankruptcy in academic and popular press and nasty political battles over tenure. These unpleasant incidents not only help no one but also obscure the fact that *both* the academic sciences *and* the humanities are facing crises of funding in an economy that values quick profit and immediate reward over a long-term investment in knowledge. In the end, neither science nor science studies benefits from a situation best summed up from *both* sides by Alan Sokal's complaint: "The targets of my critique have by now become a self-perpetuating academic subculture that typically ignores (or disdains) reasoned criticism from the outside" [Sokal, 1996].[8]

AI Skirmishes

While most scientists remain blissfully unaware of the Science Wars, they are not unaffected by them. Within AI, the tension between the self-proclaimed defenders of scientific greatness and the self-identified opponents of scientific chauvinism is worked out under the table. In particular, the sometimes tendentious clashes between classical and alternative AI often reflect arguments about science and the role of culture in it.

This can be seen most clearly in a rather unusual opinion piece that appeared several years ago in the *AI Magazine* [Hayes *et al.*, 1994] . The remarkable rhetoric of this essay in a journal more often devoted to the intricacies of extracting commercially relevant information from databases may be appreciated in this excerpt:

> Once upon a time there were two happy and healthy babies. We will call them Representation Baby (closely related to Mind Baby and Person Baby) and Science Baby (closely related to Reality Baby).
>
> These babies were so charming and inspirational that for a long time their nannies cared for them very well indeed. During this period it was generally the case that ignorance was pushed back and human dignity increased. Nannies used honest, traditional methods of baby care which had evolved during the years. Like many wise old folk, they were not always able to articulate good justifications for their methods, but they worked, and the healthy, happy babies were well growing and having lots of fun.
>
> Unfortunately, some newer nannies haven't been so careful, and the babies are in danger from their zealous ways. We will focus on two nannies who seem to be close friends and

[8] Alan Sokal happens to be a physicist complaining about science studies, but this quote works just as aptly to summarize the complaints made the other way around.

> often can be seen together - Situated Nanny (called SitNanny for short) and Radical Social Constructivist Nanny (known to her friends as RadNanny) (15).[9]

A little decoding is in order for those not intimately aware of both the AI debates and the Science Wars. "SitNanny" represents situated action, a brand of alternative AI that focuses its attention on the way in which agents are intimately related to, and cannot be understood without, their environment. "RadNanny" is the embodiment of the cultural studies of science, social constructivism being the belief that science, like every other human endeavor, is at least partially a product of sociocultural forces (the 'radical' here functions as little more than an insult, but implies that science is *purely* social, i.e. has absolutely no relationship to any outside reality).

Having broken the code, the implication of this excerpt is clear: everything in AI was going fine as long as we thought about things in terms of science and knowledge representation. Of course, this science was not always well-thought-out, but it was fundamentally good. That is, until that dastardly alternative AI came along with cultural studies in its tow and threatened nothing less than to *kill the babies*.

Note for the non-AI readers: knowledge representation can be thought of as the belief that AI agents have explicit representations of the outside world in their head, which they manipulate in order to forecast what affect their actions will have on the world.

Now any cultural critic worth his or her salt will have some choice commentary on a story in which the positive figures are all male babies living the life of leisure, and the negative figures all lower-class working women.[10] But the really interesting rhetorical move in this essay is in the alignment of the classical-alternative AI debate with the Science Wars. Classical AI, we learn, is good science. Alternative AI, while having some good ideas, is dangerous, among other reasons because it is watering down science with other ideas: "concepts from fringe neurology, sociology, ethnomethodology, and political theory; precomputational psychological theory; and God knows what else" (19). Alternative AI is particularly dangerous because it believes that agents cannot be understood without reference to their environment. Hence, it is allied with the "cult" (20) of science studies, which believes that scientists cannot be understood without reference to their sociocultural environment.

Since the majority of their audience presumably has little awareness of science studies, the authors are happy to do their part for interdisciplinary awareness by explaining what it is. They state, in a particularly nice allusion to 1950's anti-Communist hysteria, that science studies aims at nothing less than to "reject the entire fabric of Western science" (15). Science studies, we are informed, believes "that all science is arbitrary and that reality is merely a construction of a social game" (23). In the delightful tradition of the Science Wars, several quotations are taken out of context to prove that cultural critics of science believe that science is merely an expendable myth.

The statements Hayes et. al. make are simply inaccurate descriptions of science studies. In reality, science studies tends to be agnostic on such questions as the arbitrariness of science and on the nature of reality, to

[9] This excerpt cannot, however, carry the full force of the original, which contains several full-page 19th-century woodcuts displaying suffering babies and incompetent or evil nannies (labeled, for example, "The Notorious RadNanny Looking For Babies").

[10] One must presume that the authors were aware of this and did their best to raise cultural critics' hackles.

which science studies generally does not claim to have any more access than science does. When science studies *does* look into these issues it does so in a much more subtle and complex way than simply rejecting or accepting them.

But what is more important than these factual inaccuracies is that the article promotes the worst aspects of the Science Wars, since the very *tone* of the article is chosen to preclude the possibility of productive discussion. Science studies is simply dismissed as ludicrous. If uninformed scientists reading the article have not by the end concluded that science studies is an evil force allied against them, with alternative AI its unfortunate dupe, it is certainly not for lack of trying.

AI in Culture, AI as Culture

But is it really true that science studies is an enemy of AI? After all, no one disputes that AI is, among other things, a social endeavor. Its researchers are undeniably human beings who are deeply embedded in and influenced by the social traditions in which they consciously or unconsciously take part, including but by no means limited to the social traditions of AI itself. It seems that taking these facts seriously might not necessarily damage AI, but could even help AI researchers do their work better.

In this section, we will buck the trend of mutual disciplinary antagonism by exploring the potential of what Agre calls *critical technical practices* [Agre, 1997]. A critical technical practice is a way of actually doing AI which incorporates a level of reflexive awareness of the kind espoused by science studies. This may include awareness of the technical work's sociocultural context, its unconscious philosophies, or the metaphors it uses. We will look at various AI researchers who have found ideas from science studies helpful in their technical work. With this previous work as its basis, in the rest of this chapter I will explain the approach to synthesizing AI and cultural studies I am taking in this thesis.

A Short History of Critical Technical Practices

From the rather heated rhetoric of the Science Wars, you might be tempted to think that science and science studies have nothing of value to share with each other. Often, voices on the 'pro-science' side of the debate say that the cultural studies of science has no right to speak about science because only scientists have the background and ability to understand what science is about and judge it appropriately. At the same time, the 'pro-culture' side of the debate may feel that scientists neither know about nor care to ameliorate the social effects of their work.[11]

[11] The way in which these attitudes cut off communication is not infrequently illustrated to me in the flesh. For example, a cultural theorist who was once introduced to me immediately said, "So, you work in AI. How does it feel to be the instrument of global capital in the replacement of workers by machinery?" I immediately responded, "I don't know. How does it feel to be the instrument of the university in the training of the next generation of happy materialist consumers?" — not because the question was unreasonable, but because its very phrasing demonstrated that the possibility for meaningful communication had been deliberately closed off from the start. Lest cultural theorists be singled out for judgment, in my experience scientists are quite capable of similar 'conversations.'

These attitudes can only be maintained by studiously avoiding noticing the people who are *both* scientists *and* cultural critics. Gross and Levitt's influential onslaught against science studies, for example, argues that cultural critics are irresponsible and dangerous because they are ignorant of the science they criticize. This argument is made easier by counting interdisciplinarians who do *both* science *and* cultural studies as (good, responsible) scientists and not as (bad, irresponsible) cultural critics (the question of why those scientists would find it interesting or even fruitful to keep such unseemly company is left unanswered) [Gross and Levitt, 1994]. And in an exhaustive survey of every important figure in cultural studies, some of the most influential 'culturalist scientists' are left out together. A glaring omission is Richard Lewontin, whose influential books on the cultural aspects of biology are the sidelight to an illustrious career as a geneticist [Levins and Lewontin, 1985] [Lewontin *et al.*, 1984].[12]

Similarly, the hypothesis that scientists do not know or care about the effects of their work is contradicted by the work of Martha Crouch [Crouch, 1990]. Crouch is a botanist who, after many years of research, noticed that the funding of botany combined in practice with the naive faith of scientists in their own field to completely undermine the idealistic goals of plant scientists themselves. Crouch determined to help scientists such as herself achieve their own stated goals of, for example, feeding the hungry, by adding to their self-understanding through the integration of cultural studies with botany.

But, to be fair, much of the work integrating science with science studies may be invisible to both cultural critics themselves and the scientists whose form of intellectual output seems to largely be attacks on those on the other side of the great intellectual divide. This is because scientists who are actually using culturalist perspectives in their work generally address that work to their scientific subcommunity, rather than to all of science and science studies as a whole. And in work that is addressed to a technical subfield, it is usually not particularly advantageous to mention that one's ideas stem from the humanities, particularly if they come from such unseemly company as hermeneutics, feminism or Marxism.

Here, we will uncover the history of the use of culturalist perspectives within AI as a part of technical work. It turns out that within AI, the use of the humanities is not just a couple of freak accidents traceable to a few lone geniuses and / or lunatics. Rather, there is a healthy if somewhat hidden tradition of a number of generations of AI researchers who have drawn inspiration from the humanities in ways that have had substantial impact on the field as a whole. We will be interested both in finding out how cultural studies was found to be useful, and in the concrete methods various researchers have used to combine the fields.

Winograd and Flores

Terry Winograd is one of the first and certainly one of the most notorious in his usage of critical theory to analyze AI from the AI researcher's point of view. As mentioned in the review of classical AI, Winograd was a well-known researcher into the machine generation of human language.

[12] For Lewontin's roasting response to Gross and Levitt, see [Lewontin, 1995].

Upon collaboration with Fernando Flores, an economist, Winograd became interested in Heidegger. In an unexpected move, after trying to understand AI in a Heideggerian sense Winograd chose to jettison AI altogether as impossible.

In [Winograd and Flores, 1986], Winograd and Flores describe AI as fundamentally too invested in the analytic tradition to ever be able to address fundamental attributes of intelligence. In particular, they focus on the Heideggerian idea that a person in the world is always operating from a set of background prejudices that can not be finitely, hence mechanically, articulated. AI *can* solve problems that are formally specified and circumscribed, but will always fail to attain true intelligence because "[t]he essence of intelligence is to act appropriately when there is no simple pre-definition of the problem or the space of states in which to search for a solution" (98).

Prejudice here refers to things that you subconsciously believe without having justified them, as opposed to negative stereotypes of people different from yourself.

While Winograd and Flores's arguments certainly made a splash in the field, it must be honestly stated that they probably did not cause too many scientists to leave AI (and they were not intended to). The basic flaw from this perspective in the argument is that it forces AI researchers to choose between believing in Heidegger and believing in AI. One can hardly blame them if they stay with the known evil.

What is interesting to those who remain in AI, however, is Winograd and Flores's methodology for combining a philosophical perspective with AI. Winograd and Flores analyze the limitations of AI that stem from its day-to-day methodologies. When they find those constraints to exclude the possibility of truly intelligent behavior, they decide instead to start building systems in which those constraints become strengths. In other words, they decide that artificial systems necessarily have certain characteristics of rigidity and literalness, then ask themselves what sorts of social situations could be aided by a rigid, literal system. They then build a system that is an enforcer of social contracts in certain, limited situations where they feel it is important that social agreements be clearly delineated and agreed upon. Specifically, the system articulates social agreements within work settings, so that workers are aware of who has agreed to do what. This new system is designed to be useful precisely because of the things that were previously limitations! Winograd and Flores, then, use cultural studies to inform technical development by finding constraints in its methodologies, and then using those constraints so that they become strengths.

Suchman

Lucy Suchman is an anthropologist who, for a time, studied AI researchers and, in particular, the ideas of 'planning' [Suchman, 1987]. Planning is an area of AI that is, at its most broad, devoted to deciding what to do. Since this broad conception does not really help you sink your teeth into the problem, a more limited notion has been generally used in AI. This concept of planning is a type of problem-solving where an agent is given a goal to achieve in the world, and tries to imagine a set of actions that can achieve that goal, generally by using formal logic.

Suchman noticed that the ideas of planning were heavily based on largely Western notions of, among other things, route planning. She then asked herself what kind of 'planning' you would have if you used the notions of a different society. By incorporating perspectives from Micronesian society, she came up with the concept of 'situated action,' which you may remember as the butt of ridicule in Hayes et. al.'s "On Babies and Bathwater" (page 11).

Situated action's basic premise is to generate behavior on the fly according to the local situation, instead of planning far ahead of time. Although Suchman herself made no claims to technical fame, her ideas became influential among AI researchers who were working on similarly-motivated technology (see below), becoming an important component in an entire subfield AI researchers now either love or hate, but generally cannot ignore. Her methodology, in sum, is to notice the culture-boundedness of a particular metaphor ("planning") that informs technical research, then ask what perspectives a very different metaphor might bring to the field instead. The point in her work is not that Western metaphors are 'wrong' and non-Western ones are 'right,' but that new metaphors can spawn new machinery that might be interesting in different ways from the old machinery.

Chapman

David Chapman was a graduate student at MIT when together with Agre, whose work is described separately below, he developed an agent architecture that was heavily influenced by Suchman's ideas, as well as by ethnomethodology [Chapman, 1990]. This architecture is described in more detail in Chapter 2. Chapman's contribution in this history of interdisciplinary methodologies in AI is his articulation of the value of 'ideas' — as opposed to proofs or technical implementation — in technical practice.

Chapman argues that some of the most interesting papers in AI do not make technical contributions in any strict sense of the term — i.e., that the best methodology for AI is not necessarily that of empirical natural science. "[Some of the best] papers prove no theorems, report no experiments, offer no testable scientific theories, propose technologies only in the most abstract terms, and make no arguments that would satisfy a serious philosopher.... [Instead, t]hese papers have been influential because they show us powerful ways of thinking about the central issues in AI" (214). Suchman's anthropological work in AI is a living example in Chapman's work of such an influential idea.

Agre

Of all AI researchers, Agre has probably done the most extensive and explicit integration of critical viewpoints with AI technology. In his thesis, for example, Agre integrates ethnomethodology with more straightforward AI techniques [Agre, 1988]. He uses ideas from ethnomethodology both to suggest what problems are interesting to work on (routine behavior, instead of expert problem-solving) and to suggest technical solutions (deictic, or subjective representation instead of objective representation).

Together with Chapman, Agre uses a philosophical approach influenced by Winograd's Heideggerian analysis of AI, but based more primarily on the work of such ethnomethodologists as Suchman and Garfinkel, to develop not only a new methodology for building agents, but also a new understanding of what it means to be an agent in the world that goes beyond views of life as consisting of rational problem-solving.

> The world of everyday life... is not a problem or a series of problems. Acting in the world is an ongoing process conducted in an evolving web of opportunities to engage in various activities and contingencies that arise in the course of doing so.... The futility of trying to control the world is, we think, reflected in the growing complexity of plan executives. Perhaps it is better to view an agent as *participating* in the flow of events. An embodied agent must *lead a life*, not *solve problems* ([Agre and Chapman, 1990], 20).

This re-understanding of the notion of agent has been an important intellectual strand in alternative AI's reconceptualization of agent subjectivity.

In recent work, Agre has distilled his approach to combining philosophy, critical perspectives, and concrete technical work into an articulated methodology for critical technical practices per se. Agre sees critical reflection as an indispensable tool in technical work itself, because it helps technical researchers to understand in a deep sense what technical impasses are trying to tell them. He sums up his humanistic approach to AI with these postulates:

> 1. AI ideas have their genealogical roots in philosophical ideas. 2. AI research programs attempt to work out and develop the philosophical systems they inherit. 3. AI research regularly encounters difficulties and impasses that derive from internal tensions in the underlying philosophical systems. 4. These difficulties and impasses should be embraced as particularly informative clues about the nature and consequences of the philosophical tensions that generate them. 5. Analysis of these clues must proceed outside the bounds of strictly technical research, but they can result in both new technical agendas and in revised understandings of technical research itself. [Agre, 1995]

Humanists will recognize Agre's methodology as a kind of hermeneutics, i.e. a process of interpretation that goes beyond surface appearances to discover deeper meanings. For Agre, purely technical research is the surface manifestation of deeper philosophical systems. While it is certainly possible for technical traditions to proceed without being aware of their philosophical bases, technical impasses provide clues that, when properly interpreted, can reveal the philosophical tensions that lead to them. If these philosophical difficulties are ignored, chances are that technical impasses will proliferate and remain unresolved. If, however, they are acknowledged, they can become the basis for a new and richer technical understanding.

In [Agre, 1997], Agre develops a methodology for integrating AI and the critical tradition through the use of deconstruction. Deconstruction is a technique developed by philosopher Jacques Derrida for analyzing texts in order to bring out inherent contradictions hidden in them [Derrida, 1976] [Culler, 1982]. Agre's methodology involves the following steps:

Dear humanists, forgive me for this reductive explanation, but *you* try explaining deconstruction to engineers in one sentence or less.

1. Find a metaphor that underlies a particular technical subfield. An example of such a metaphor is the notion of disembodiment that underlies classical AI.

2. Think of a metaphor that is the *opposite* of this metaphor. The opposite of disembodied agents would be agents that are fundamentally embodied.

3. Build technology that is based on this opposite metaphor. Embodied agents are an essential component of Rod Brooks's groundbreaking work, which is described in more detail in Chapter 2. This technology will inevitably have both new constraints and new possibilities when compared to the old technology.

In Agre's work, metaphorical analysis can become the basis for widening our perspective on the space of possible technologies.

Varela, Thompson, and Rosch

Francisco Varela, Evan Thompson, and Eleanor Rosch do not combine AI with cultural studies. Varela is a well-known cognitive scientist (a sister discipline of AI); Thompson and Rosch are philosophers. Nevertheless, their work is closely related to syntheses of AI and the humanities and deserves to be addressed along with them.

In [Varela *et al.*, 1991], Varela, Thompson and Rosch integrate cognitive science with Buddhism, particularly in the Madhyamika tradition. They do this by connecting cognitive science as the science of cognition with Buddhist meditation as a discipline of experience. Current trends in cognitive science tend to make a split between cognition and consciousness, to the point that some cognitive scientists call consciousness a mere illusion. Instead, Varela et. al. connect cognition and experience so cognitive scientists might have some idea of what their work has to do with what it means to be an actual, living, breathing human being.

Varela, Thompson, and Rosch stress that cognitive science — being the study of the mind — should be connected to our actual day-to-day experience of what it means to have a mind. What they mean here by experience is not simple existence per se but a deep and careful examination of what that existence is like and means. They believe that your work should not deny or push aside your experience as a being in the world. Instead, that experience should be connected to and affirmed in your work. In this way, they connect with cultural critics of science like Donna Haraway and cultural theorists like Gilles Deleuze and Félix Guattari, who stress the importance of personal experience as a component of disciplinary knowledge [Haraway, 1990b] [Deleuze and Guattari, 1987].

One of the tensions that has to be resolved in any work that combines science with non-scientific disciplines (of which Buddhism is certainly one!) is the differential valuation of objectivity. Generally speaking, the humanities tend to value subjective knowledge, whereas the sciences and engineering tend to prefer results that are objective. The notion of 'objectivity' is itself a can of worms, but we can work here with a preliminary understanding of objectivity as knowledge that is independent of anyone's individual, personal experiences. Since Varela, Thompson and Rosch want to connect cognitive science as science with individual human experience, they confront this problem of subjectivity versus objectivity head-on.

Interestingly, they do this by redefining what objectivity means with respect to subjective experiences. You cannot truly claim to be objective, they say, if you ignore your most obvious evidence of some phenomenon, i.e. your personal experience of it. This is particularly true when one is studying cognition —- in this frame of thought, any self-respecting study of the mind should be capable of addressing the experience of having one!

Given that one of the things cognitive scientists (and, by extension, AI researchers) are or should be interested in is subjective experience, Varela, Thompson, and Rosch abandon the focus on objectivity per se. But this does not lead to the long-feared nihilistic abandonment of any kind of judgments of knowledge — black is white, up is down, whatever I say goes, etc. Rather, they stress that Buddhist traditions have *disciplined* ways of thinking about that experience. The problem, they say, is not with subjectivity, but with being undisciplined. The goal, then, is being able to generate a kind of cognitive science that is subjective *without being arbitrary*.

Summary: Perspectives on Integrating AI and the Humanities

Generally, each of these researchers is interested in AI because of a fascination with the nature of human experience in the world. This interest naturally leads them to the humanities, which have dealt with questions of subjective human experience for hundreds of years. These researchers have found various ways to integrate this humanist experience with the science and engineering practices of AI. With respect to the issue of integrating AI and cultural studies that is pursued in this thesis, we can sum up their perspectives as follows:

- Winograd and Flores contrast existentialist philosophy with the analytic, rationalist philosophy that underlies much AI research. They use the differences between these approaches to understand the constraints that are inherent in AI methodology. They then develop new technology that, instead of being limited by these constraints, takes advantage of them.

- Suchman analyzes current AI practices to uncover the metaphors that underly them. These metaphors turn out to be specific to Western culture. She then asks what technology would be like if it were based on metaphors from a different culture.

- Chapman implements technology that is deeply informed by, among other things, the newly-identified metaphors of Suchman. He defends the concept that, though technology is well and good, fundamental *ideas* that are not testable in a scientific or mathematical sense are equally valuable to AI.

- Agre understands technical work as reflecting deep philosophical tensions. From this point of view, technical problems *are* philosophical problems. This means that the best progress can be made in AI by thinking simultaneously at the technical and at the philosophical levels.

- Varela, Thompson, and Rosch connect the science of human cognition with the subjective experience of human existence. They introduce, flesh out, and defend the idea that subjective does not necessarily mean arbitrary.

Each of these themes will be taken up in the work that follows.

Methodology: Subjective Technologies

Note to the technically trained: this section is philosophical and personal; its style may feel unfamiliar and uncomfortable for you. It inherits its style more from the traditions of cultural studies than technical work. I recommend trying to read it with a poetic rather than a technical frame of mind. If you do so, you may find that it not only lays out important foundations for the arguments that are to follow, but also betrays many secrets to the origins of my technical work that you, too, may find helpful in yours.
In this, you may find helpful the perspective of László Mérő: "My native language is rationality; my everyday logic cannot accept conclusions that contradict scientific results. Yet at the same time I clearly feel that there are many fields that slip out of the present range of science — and I do not deem them unworthy of reflection." ([Mérő, 1990],52)

The approach taken in this thesis follows Varela, Thompson, and Rosch in asserting that subjective experience, which goes to the heart of what it means to humans to be alive in the world, should be an important component of AI research. I believe that one of the major limitations of current AI research — the generation of agents that are smart, useful, profitable, but not convincingly alive — stems from the traditions AI inherits from science and engineering. These traditions tend to discount subjective experience as unreliable; the experience of consciousness, in this tradition, is an illusion overlaying the actual, purely mechanistic workings of our biological silicon. It seems to me no wonder that, if consciousness and the experience of being alive are left out of the methods of AI, the agents we build based on these methods come across as shallow, stimulus-response automatons.

In the reduction of subjective experience to mechanistic explanations, AI is by no means alone. AI is part of a broader set of Western cultural traditions, such as positivist psychiatry and scientific management, which tend to devalue deep, psychological, individual, and subjective explanations in favor of broad, shallow, general, and empirically verifiable models of the human. I do not deny that these theories have their use; but I fear that, if taken as the *only* model for truth, they leave out important parts of human experience that should not be neglected. I take this as a moral stance, but you do not need to accept this position to see and worry about the symptom of their neglect in AI: the development of agents that are debilitatingly handicapped by what could reasonably accurately, if metaphorically, be termed autism.

This belief that science should be understood as one knowledge tradition among others does not imply the rejection of science; it merely places science in the context of other, potentially — but not always actually — equally valid ways of knowing. In fact, many if not most scientists themselves understand that science cannot provide all the answers to questions that are important to human beings. This means that, as long as AI attempts to remain purely scientific, it may be leaving out things that are essential to being human.

In *Ways of Thinking: The Limits of Rational Thought and Artificial Intelligence*, for example, cognitive scientist László Mérő, while affirming his own scientific stance, comes to the disappointing conclusion that a scientific AI will inevitably fall short of true intelligence.

> In his book *Mental Models* Johnson-Laird says, 'Of course there may be aspects of spirituality, morality, and imagina-

> tion, that cannot be modeled in computer programs. But these faculties will remain forever inexplicable. Any scientific theory of the mind has to treat it as an automaton.' By that attitude science may turn a deaf ear to learning about a lot of interesting and existing things forever, but it cannot do otherwise: radically different reference systems cannot be mixed. (228-229)

But while the integration of science and the humanities (or art [Penny, 1997b] or theology [Foerst, 1998] [Foerst, 1996]) is by no means a straightforward affair, the work already undertaken in this direction by researchers in AI and other traditionally scientific disciplines suggests that Mérő's pessimism does not need to be warranted. We *do* have hope of creating a kind of AI that can mix these 'radically different reference systems' to create something like a 'subjectivist' craft tradition for technology. Such a practice can address subjective experience while simultaneously respecting its inheritances from scientific traditions. I term these perhaps heterogeneous ways of building technology that include and respect subjective experience 'subjective technologies.' This thesis is one example of a path to subjective technology, achieved through the synthesis of AI and cultural studies, but it is by no means the only possible one.

Because of the great differences between AI and cultural studies, it is inevitable that a synthesis of them will include things unfamiliar to each discipline, and leaves out things that each discipline values. In my approach to this synthesis, I have tried to select what is to be removed and what is to be retained by maintaining two basic principles, one from AI and one from cultural studies: (1) faith in the basic value of concrete technical implementation in complementing more philosophical work, including the belief that the constraints of implementation can reveal knowledge that is difficult to derive from abstract thought; (2) respect for the complexity and richness of human and animal existence in the world, which all of our limited, human ways of knowing, both rational and nonrational, both technical and intuitive, cannot exhaust.

"[T]he *interdisciplinarity* which is today held up as a prime value in research cannot be accomplished by the simple confrontation of specialist branches of knowledge. Interdisciplinarity is not the calm of an easy security: it begins *effectively*... when the solidarity of the old disciplines breaks down... in the interests of a new object and a new language, neither of which has a place in the field of the sciences that were to be brought together." ([Barthes, 1984], 169)

The Anti-Boxological Manifesto

The methodologies I use here inherit many aspects from the previous work described above. Following Winograd and Flores, I analyze the constraints that AI imposes upon itself through its use of analytic methodologies. Following Suchman, I uncover metaphors that inform current technology, and search for new metaphors that can fundamentally alter that technology. Following Chapman, I provide not just a particular technology of AI but a way of thinking about how AI can be done. Following Agre, I pursue technical and philosophical arguments as two sides of a single coin, finding that each side can inform and improve the other.

The additions I make to these approaches are based on a broad analysis of attempts to limit or circumscribe human experience. I believe that the major way in which AI and similar sciences unintentionally drain the human life out of their objects of study is through what agent researchers Petta and Trappl satirize as 'boxology:' the desire to understand phenomena in the world as tidy black boxes with limited interaction [Petta and Trappl, 1997]. In order to maintain the comfortable illusion that these black boxes sum up all that is important of experience, boxologists are forced to ignore or devalue whatever does not fall into the neat categories that are set up in their sstem. The result is a view of life that is attractively simple, but with glaring gaps, particularly in places where individual human experience contradicts the established wisdom the categories represent.

The predominant contribution to this tradition of humanistic AI which this thesis tries to make is the development of an approach to AI that is, at all levels, fundamentally anti-boxological. At each level, this is done through a contextualizing approach:

- *At the disciplinary level*, rather than observing a strict division of technical work and culture, I synthesize engineering approaches with cultural insights.
- *At the methodological level*, rather than designing an agent as an independent, autonomous being, I place it in the sociocultural context of its creators and the people who interact with it.
- *At the technical level*, rather than dividing agents up into more or less independent parts, I explicitly place the parts of the agent in relation to each other through the use of mediating transitions.

At all levels, my approach is based on this heuristic: "that there is no such thing as relatively independent spheres or circuits" ([Deleuze and Guattari, 1977], 4). My approach may feel unusual to technical workers because it is heavily metaphorical; I find metaphorical connections immensely helpful in casting unexpected light on technical problems. I therefore include in the mix anything that is helpful, integrating deep technical knowledge with metaphorical analysis, the reading of machines ([Mahoney, 1980]), hermeneutics, theory of narrative, philosophy of science, psychology, animation, medicine, critiques of industrialization, and, in the happy phrasing of Hayes and friends, "God knows what else." The goal is not to observe disciplinary boundaries — or to transgress them for the sake of it — but to bring together multiple perspectives that are pertinent to answering the question, "What are the limitations in the way AI currently understands human experience, and how can those limitations be addressed in new technology?"

Preview of Thesis Yet to Come

This phrasing of the fundamental question of the thesis may be a little too general for your tastes. In the next chapter, we will begin focusing on a detailed technical question: how to integrate many complex behaviors in an agent without degrading its overall quality of activity. The general goal of the thesis is to integrate engineering with humanistic perspectives; the concrete goal is to find technical solutions for behavioral degeneration by understanding its origin in the methodologies for agent interpretation and construction that are part of AI's scientific inheritance.

I will approach this goal in several steps. In Chapter 2, I will review current AI methodologies for synthesizing behavior, and uncover an inevitable limitation in its current approach. In Chapter 3, I will deepen this understanding by comparing AI approaches to agenthood with the methods of positivist psychiatry and scientific management; these black-boxing, objective approaches to human experience will be contrasted with contextualizing, subjective approaches critics of them have devised. In Intermezzo I, I will briefly introduce the "Industrial Graveyard," an implemented virtual environment that illustrates these objective and subjective approaches to agents, and forms the testbed for the technology developed in the thesis.

I use this notion of objective and subjective approaches to agents in order to develop a 'subjectivist' extension to alternative AI, termed socially situated AI, in Chapter 4. I use this approach to redefine the problems of behavioral disintegration in terms of audience *perception* of disintegration, and develop concrete technology to address it in Chapter 5. This involves the redefinition of behaviors as communicating *signifiers*, the development of *transitions* to synthesize behaviors, and the use of *meta-level controls* to implement transitions.

It turns out, however, that the approach of Chapter 5 is in practice too limited. Basically, it inherits an engineering perspective on the notion of audience perception that turns out to be inadequate in practice. In Intermezzo II, I take a brief detour into animation to find out how *animators* create the perception of authentically living beings. I combine this perspective with narrative psychology in Chapter 6 to form a new theory of intentional behavior based on the user's construction of narrative explanations. This 'narrative intentionality' forms the core of my developed agent architecture, the Expressivator, which is presented in its full glory in Chapter 7. With the cultural analysis and technical development of autonomous agents under our belt, Chapter 8 will return to the themes of the introduction, laying out how the work done here could form a part of a future integrated scientific-humanistic AI.

A Few Remarks on Format

This thesis is interdisciplinary between two fields that share little in their background knowledge or preferred rhetorical forms. Nevertheless, the work done here is not some AI work plus some cultural studies work; it is a single piece of work that has an AI face, a cultural studies face, and a large body in between.

The format of this thesis is intended to make comprehension of this undisciplined mass of knowledge as painless to the disciplinary reader as possible. The full body of the text is written in an attempt to be understandable to both the technically and the humanistically trained. However, the inclusion of all background knowledge that one or the other side may be missing would hopelessly balloon this thesis out of proportion and out of comprehensibility. When particular background knowledge is essential for just one discipline or the other to be able to make sense of the argument, that knowledge generally appears in sidebars to the text. Occasional sections (most notably the related work section in Chapter 5) lean heavily towards one side or the other. My hope is, however, that, for most of the thesis, no matter what your background, you will be able to negotiate a complete path through it, and find something in that path that is useful to you.

Chapter 2

Schizophrenia in Agents: Technical Background

The premise of this work is that there is something deeply missing from AI, or, more specifically, from the current dominant ways of building artificial agents. This uncomfortable intuition has been with me as an AI researcher for a long time, perhaps from the start, although for most of that time I was not able to articulate it clearly. Artificial agents seem to be lacking a primeval awareness, a coherence of action over time, something one might, for lack of a better metaphor, term 'soul.'

Roboticist Rodney Brooks expresses this worry eloquently:

> Perhaps it is the case that all the approaches to building intelligent systems are just completely off-base, and are doomed to fail. Why should we worry that this is so? Well, certainly it is the case that all biological systems.... [b]ehave in a way which just simply seems *life-like* in a way that our robots never do.
>
> Perhaps we have all missed some organizing principle of biological systems, or some general truth about them. Perhaps there is a way of looking at biological systems which will illuminate an inherent necessity in some aspect of the interactions of their parts that is completely missing from our artificial systems.... [P]erhaps at this point we simply do not *get it*, and... there is some fundamental change necessary in our thinking in order that we might build artificial systems that have the levels of intelligence, emotional interactions, long term stability and autonomy, and general robustness that we might expect of biological systems... [P]erhaps we are currently missing the *juice* of life. ([Brooks, 1997], 299-300)

This lack of 'aliveness' is not just a fuzzy intuition; it has its technical manifestations. One way in which this lack is expressed is in the difficulty of creating complex artificial creatures. A popular way of building these creatures in the alternative AI tradition is by composing behaviors. We have well-developed techniques for building behaviors which, by themselves, are clear, expressive and giving off the appearance of

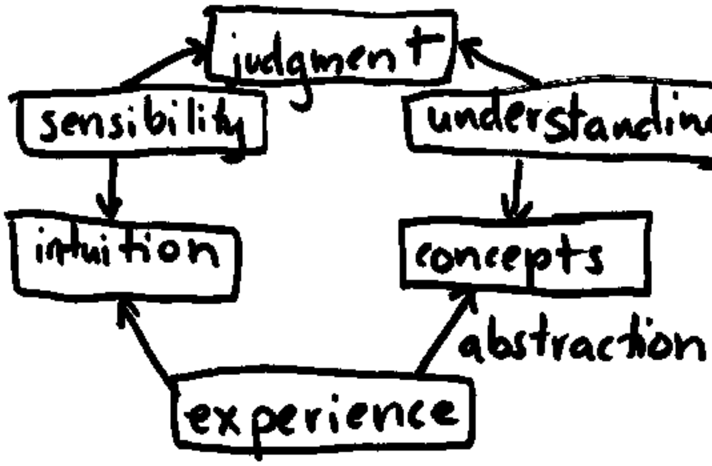


FIGURE 2.1: An agent structure inspired by Kant

life. The problem is that, as we try to combine more and more of these behaviors, the agent's overall activity gradually falls apart. If only a few behaviors are involved, the programmer can generally manage this disintegration. But when many behaviors are involved, their interactions are too complicated to be easily managed by hand.

The end effect of this difficulty is that, practically speaking, many complex behaviors simply cannot be adequately integrated. Instead, the agent tends to jump around from behavior to behavior, abruptly switching from one internally coherent behavior to another, its final activity a crazy quilt of actions with no coherent thread. These creatures, while perhaps intelligent in a formal sense, do not appear to have the coherence of behavior over time that we impute to living creatures. I term this overall incoherence *schizophrenia*, for reasons that will be thoroughly discussed later in this chapter.

In this chapter, we will examine this problem in the development of autonomous agents in detail. I give an overview of alternative approaches to agent construction, and then identify particular difficulties that tend to come up in synthesizing these agents. We will look at the construction of autonomous agents in depth to understand why schizophrenia happens. It turns out that the problem of schizophrenia is deeply connected with the way we think about building agents per se. Understanding this connection will provide the foundation for rethinking agent construction and addressing schizophrenia in the remainder of the thesis.

How to Build Yourself an Agent

You should note that this way of conceptualizing humanistic traditions, while hopefully helping with the notion of agent construction, simultaneously does a grotesque violence to them, of a form which will become clearer in Chapter 3.

It can sometimes be difficult for non-technical readers to imagine what exactly the parts of an agent might be, or how they could be connected to build a complete agent. Prior to delving into the guts of doing this from a technical point of view, I have taken the liberty of building two diagrams that show how an agent designed by a humanist might look (see Figures 2.1 and 2.2 — please take these with a liberal grain of salt). I

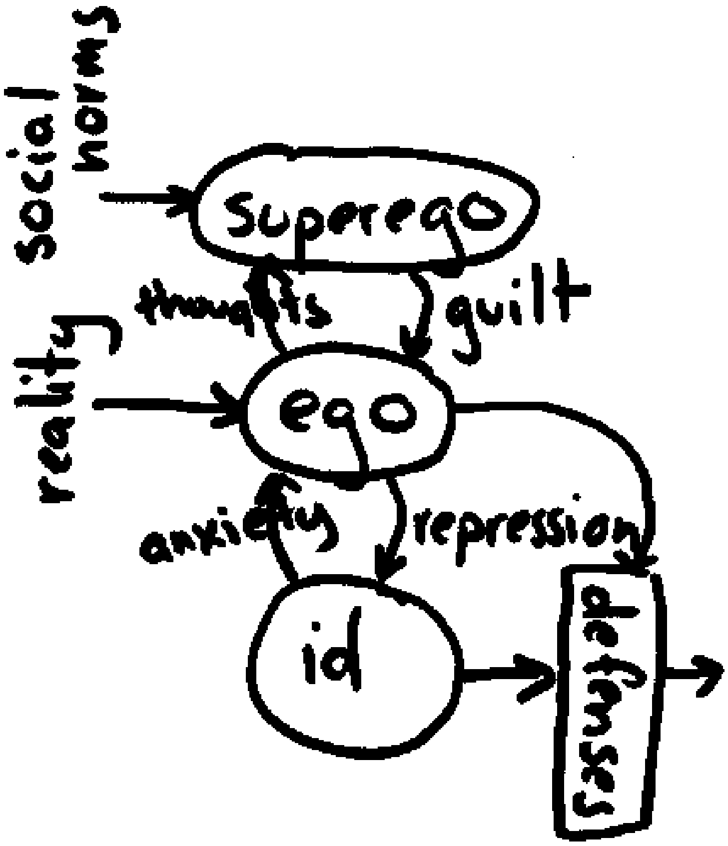


FIGURE 2.2: An agent structure inspired by Freud

heartily encourage the reader to design his or her own agent in the empty box provided for this purpose (Figure 2.3). Try to imagine what the various parts would be and how they might be interconnected. Please remember that anything not specified directly will not exist; do not exceed the boundaries of the box.

Pretty difficult, huh? My guess is that most people working from a humanistic tradition will quickly throw in the towel, since subjectivity is not something that can be simply diagrammed out on a piece of paper. AI researchers have no such luxury. The only way to build something is to specify it exactly. This means an essential part of agent construction is (a) deciding what the parts of an agent are and (b) deciding how the parts of an agent should be interconnected.

Until recently, the focus of the classical AI tradition has largely been on answering the first question. Through the mid-80's, classical AI research projects tended to focus on the development of isolated components for agents. Typically these components included natural language understanding systems, vision systems, memory modules, or planners. The final integration of agents into a complete, embodied, fully functional system was often deferred until the parts were sufficiently stable, which generally meant at some point in the distant future.

When some brave souls did attempt such integration, results were

FIGURE 2.3: Draw your own agent here!

often disappointing, particularly in the robotic domain. Shakey [Nilsson, 1984], a beloved yet accurately named robot built in the late 60's and early 70's at the Stanford Research Institute, was one of the first attempts at building a complete agent. Shakey would go through cycles of sensing the environment, building an internal model of the outside world, deciding what to do, and doing it. Even in a carefully engineered environment, each of these cycles could take upwards of an hour. Splitting Shakey up into these sense-map-plan-act stages introduced computational bottlenecks that drastically affected its ability to react to a potentially changing environment.

More fundamentally, focusing on components and their subsequent specialization in research ghettos means that there are no forums to address their interrelationships. Systems are built with different logics, different input and output interfaces, and different assumptions about what the other systems will or can do. The temptation to leave out parts that are particularly difficult or ill-defined is strong, and there is no particular reason to resist it (or was, prior to the arguments of alternative AI). A somewhat crude but effective characterization of classical AI for humanists in this light is as the separate rationalization of part processes, with the eventual coordination of these processes into a coherent whole infinitely deferred.

Alternative AI defines itself in opposition to this approach as attempting to construct complete agents, from the ground up. The focus in these projects is often on building a complete agent first, then gradually improving its capabilities in a succession of more and more competent agents. A necessary and recurring preoccupation for these agent-builders, then, is the question of how the various pieces of an agent can be appropriately combined to form an at least semi-coherent agent.

Note that alternative AI does not have a monopoly on complete agent construction. A nice example of a classical agent is Homer [Vere and Bickmore, 1990], a virtual submarine that lives its own (rather dull) life under the sea, while taking orders from its buddies, Tim and Steve. It is also clear that many alternative AI projects also simply focus on small components. The point here is alternative AI's explicit commitment to and interest in integrated systems.

Typical agent parts:

Classical AI	Alternative AI
Perception	Object Avoidance
Modeling	Wandering
Planning	Wall Following
Execution	Picking Up Objects
Natural Language	Recharging Batteries

In the next sections, we will look at some of these projects in detail to identify alternative AI perspectives on integrated agent construction. I will focus both on the *concept* of agent used and on the agent *construction techniques*. Somewhat unsurprisingly, it turns out that these two aspects are inescapably intertwined.

As there are now enough proposed agent architectures to make several years of bed-time reading for an architecture junkie like myself, I have limited myself here to a smattering of architectures for which reasonably substantial agents have already been implemented. This is not intended to be a comprehensive coverage of behavior-based architectures, but to give a flavor of the type and range of architectures that fit under this umbrella term. In addition, in a perhaps vain attempt to not lose non-technical readers, I have kept the description of agent architectures following rather high-level, at the cost of doing some violence to the details of how each architecture works. For more general coverage, I suggest [Maes, 1990a], [Steels and Brooks, 1995], [Laird, 1991], or [Tyrell, 1993]. For more details on each architecture described here, please refer to the suggestions in each section.

Terminology

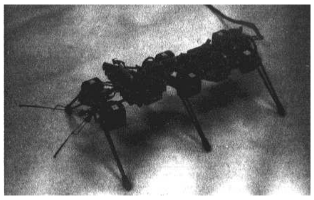
A few terms which are familiar to humanists in their colloquial sense will here be used in a technical sense. I will therefore briefly review the technical meanings of the most pertinent terms so that humanists are not immediately derailed.

- *Behavior* — A 'behavior' is a reified piece of activity in which an agent engages, for example 'sleep' or 'eat.' In colloquial English an agent behaves in various ways; in technical AIese, an agent has various behaviors.
- *The World* — When AI researchers speak of 'the world,' they mean the environment in which the agent is situated (not the Earth, for example). 'The world' is in contrast to 'the mind.'
- *Action* — An 'action' is an agent's most primitive unit of activity in the world. For typical artificial agents, actions will include things like picking up objects, rolling around, or moving arms and legs.
- *Function* — A 'function' is a reified ability which the agent has, which is often embodied in its own piece of code. Functions include things like being able to speak English, being able to see, or being able to reason about the consequences of actions.
- *Goal* — A 'goal' is a token which represents at a high level something which the agent is trying to achieve. Generally speaking, a goal is represented as a state of the world which the agent would

like to see happen: for example, that the car is parked without denting anything.

Subsumption Architecture

"Real biological systems are not rational agents that take inputs, compute logically, and produce outputs. They are a mess of many mechanisms working in various ways, out of which emerges the behavior that we observe and rationalize." ([Brooks, 1995], 52)



Brooks's Genghis

Rodney Brooks is one of the first, and certainly one of the most vocal, proponents of basing AI research on integrated agents from the start [Brooks, 1986a] [Brooks, 1986b] [Brooks, 1990] [Brooks, 1991b]. Brooks complains that previous approaches to building intelligence have often focused purely on building a "brain-in-a-box," i.e., defining agents only as information processors, without regard to their physicality. Automatic perception of the agent's physical environment, for example, has often been ignored, in favor of spoon-feeding agents human-designed descriptions of the world. Input and output being filtered through a human allows the researcher to showcase the intelligence of their subsystem, while avoiding the pesky little details of perception — which, it turns out, is extremely difficult.[1]

In contrast to this Cartesian, abstract subjectivity, Brooks sees agents as fundamentally physical and embodied. Rather than defining an agent in terms of abstract problem-solving — the chess-playing idiot savant – he thinks of it as behaving in a physical environment. The model for agenthood is inspired by biology and neurology ("Elephants don't play chess" [Brooks, 1990]), rather than human psychology. The prototypical Brooksian agent of the late 80's and early 90's[2] is the "Robot Insect." These insects are extremely limited in intelligence in comparison with traditional AI agents, but unlike these agents, they can walk rapidly around an office environment without killing anyone.

"True intelligence requires a vast repertoire of background capabilities, experience and knowledge (however these terms may be defined). Such a system can not be designed and built as a single amorphous lump. It must have components.... But true intelligence is such a complex thing that one can not expect the parts to be built separately, put together and have the whole thing work. We are in such a state of ignorance that it is unlikely we could make the right functional decomposition now. Instead we must develop a way of incrementally building intelligence." ([Brooks, 1986b],5)

Brooks's goal is to build complete agents that can function in a physical environment; he is less interested in the development of components than in the creation of complete agents, no matter how simple. As a consequence, he has problems with the way classical AI divides up its agents. He considers functional decomposition — the division of an agent into its hypothesized internal functions — to be an act of supreme intellectual arrogance. The claim is that we know so little about how agents are or should be constructed, that we will inevitably make bad choices and spend years of work on an extensive and well-designed module that will then simply be thrown away.

Since we have no way of knowing what the "proper" internal structure of an agent is, Brooks suggests that we should design an agent in terms of things we can see — its behavior. Each internally-defined agent behavior should directly connect perception of the world with action, causing humanly perceptible behavior. Just as evolution gradually builds up more and more complex animals, Brooks suggests creating more and more complex agents by adding new behaviors on top of old ones. The result is a hierarchy of behaviors, each of which is always active.

Brooks terms the typical classical AI method of dividing up agents 'horizontal decomposition' (Figure 2.4), because information from the

[1] ... though not un-tried, particularly by cyberneticists.

[2] More recently, Brooks has been building a humanoid robot that models early infant development [Brooks and Stein, 1993].

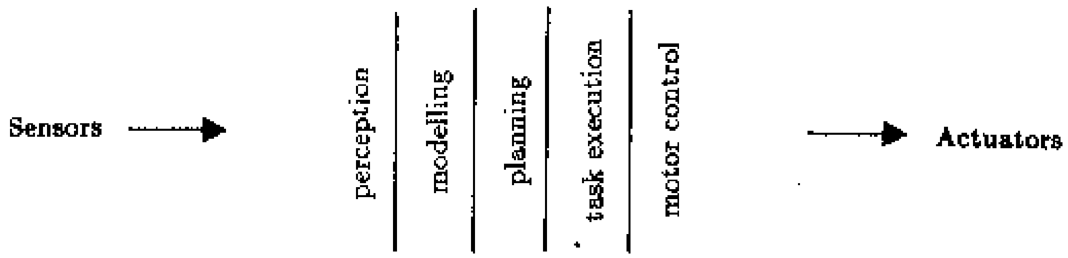


FIGURE 2.4: Horizontal Decomposition

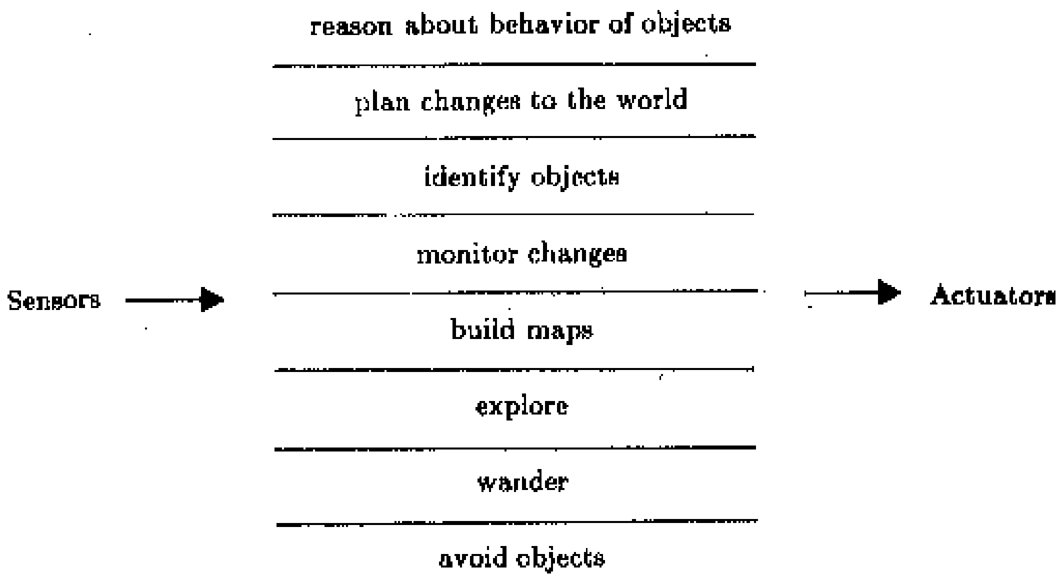


FIGURE 2.5: Vertical Decomposition

environment needs to flow through all the parts of the agent before affecting its externally observable behavior. His own system he terms 'vertical decomposition' (Figure 2.5), in which every designed module forms a direct link between external environmental input and observable behavior. In this system, the 'parts' of an agent are behaviors, each of which connects perception to action, i.e. whose effects are directly observable by its builder.

Behaviors, in this scheme, are built separately. Behaviors are not aware of each other; each is designed as a self-contained unit. Communication between behaviors is possible, though limited, but most communication occurs by observing the results of other behaviors' actions in the world. This means behaviors are very loosely coordinated. Behaviors are thought of as many self-contained parts, only locally interacting, an idea Brooks and many other alternativists inherit from Marvin Minsky [Minsky, 1988].

Behaviors are combined through a process of layering, wherein all behaviors function simultaneously. Higher-level behaviors can subsume lower-level behaviors by blocking their output, or by providing them with false input. Behaviors that do not subsume each other can influence one another using a 'hormonal' system, inspired by Maes (page 33) which provides a kind of global state [Brooks, 1991a]. Behaviors can release 'hormones' which then may trigger other behaviors to be active. Often times, conflicts between behaviors are avoided by having them only be active under particular conditions, so that they are not likely to engage in action at the same time.

Subsumption Architecture Agent Design Strategy

1. Decide what the agent should do.
2. Decompose this into behaviors in a hierarchy from simple to complex.
3. Start building behaviors from the bottom up, starting with simplest.
4. Once the simplest behavior works, design the next behavior on top of that.
5. Continue until all behaviors function.

Pengi

"The world of everyday life... is not a problem or a series of problems. Acting in the world is an ongoing process conducted in an evolving web of opportunities to engage in various activities and contingencies that arise in the course of doing so. Most of what you do you already know how to do, and most of the rest you work out as you go along. The futility of trying to control the world is, we think, reflected in the growing complexity of plan executives. Perhaps it is better to view an agent as *participating* in the flow of events. An embodied agent must *lead a life*, not *solve problems*." ([Agre and Chapman, 1990], 20)

One of the vital subfields of AI during the last 20 to 30 years is 'planning,' i.e. the selection of actions by an artificial agent in order to achieve its goals in the world. Prior to the mid-80's, planning algorithms typically had 3 parts: perception, plan-building, and execution. The perception phase (often short-circuited by the spoon-feeding methodology mentioned above) was used to build an internal model of the outside world. Plan-building took up the bulk of the effort, and generally consisted of mentally trying out all possible actions in the model of the world to try to find a sequence of actions that would cause the given goal to be achieved. Execution came after the fact and consisted of actually doing each step in the decided-on plan. Assuming that the planner was able to take into account every contingency, and that the executor could accurately do the actions given to it, this worked correctly even for complex goals.

This approach to agent construction places most of the burden of agent activity on reasoning about and manipulating a model of the world. Most of the agent-building effort is spent on thinking about the world, and very little on perceiving and acting. A lot of effort goes into considering contingencies and expecting the worst from a hostile environment. In the mid-80's, Phil Agre and David Chapman developed an agent, Pengi, based on a radically different model of agenthood [Agre and Chapman, 1990].

Agre and Chapman understand agents not as thinkers in a hostile world, but as doers situated in a usually benign environment. The agent spends most of its time in routine behavior, not in the planning out of details of action. Most of the agent's behavior is more or less automatic; variation and improvization happen as the agent responds routinely to a changing environment, rather than from the agent's flexibility in deciding complex sequences of actions.

Fundamentally, Agre and Chapman base their agent structure on the belief, heavily influenced by Lucy Suchman's description of situated action [Suchman, 1987], that intelligence should be understood in terms of interaction between an agent and environment, rather than in terms of the manipulations of an agent of a hostile world. For them, the challenge is to understand how routine behavior can arise and adapt to a changing environment, rather than how the system can anticipate and plan for every possible contingency.

Pengi

Given this interaction-oriented outlook, the separation of perception, planning, and execution no longer make sense. All parts of the agent should be integrated and tightly coupled with sensing and action. For Agre and Chapman, the parts of the agent are based on simple routines in which the agent should engage when placed in a particular environment. These routines are decomposed into actions, with the rationale for each action analyzed. The rationale for actions is then reduced to conditions in the environment that the agent can sense. An agent, then, consists of physical actions that are cued by sensed conditions.

In order to maintain tight coupling with the environment, agents no longer engage in a long-term perceive - think - act cycle. Rather, at every time step the agent must choose an action to take immediately based on conditions in the world. The routines the designer chose may or may not happen, since the actions are continuously redecided and in the middle of executing one "plan" actions from other plans may make more sense. The problem of actions conflicting is avoided by specifying enough conditions for each that there is only one 'right' action. It's not totally clear how this solution would work for an agent with many high-level routines, not all of which can be decided based on perceivable things in the world (Maes's architecture, described next, is in part a reaction to this).

Agre and Chapman Agent Design Strategy

1. Examine the agent or desired activity to find typical 'routines' one would engage in (often using ethnographic techniques).
2. Decompose these routines into actions. Determine rationale for each action.
3. For each action, find conditions in the world that should trigger that action according to its rationale.
4. For actions that are triggered at the same time, find additional conditions to let you choose between them.

Agent Network Architecture

Agre and Chapman's architecture has the advantages of being adaptive and reactive to changes in the environment. Pengi is an improviser who sometimes makes mistakes, but can go with the flow to generally come out on top. Pengi is fundamentally the reflection of a theory of human action, and is not intended as the peak of technological competence. You might like Pengi very much, but you probably don't want it to be running the US nuclear warhead control system.

"Given an agent that has multiple time-varying goals, a repertoire of actions that can be performed..., and specific sensor data, what actions should this agent take next so as to optimize the achievement of its goals?" ([Maes, 1993 1994],146)

For Pattie Maes, the functionality of agents is more important than theories of human agenthood. While the reactivity that comes from situated approaches is important, she is not wedded to Chapman and Agre's idea that agents are or should be fundamentally improvisers. Maes's agent definition is basically technical and functional, rather than psychological or biological; her examples of agents include planetary explorers, shop schedulers, and autonomous vacuum cleaners. As a consequence, for Maes the important thing is getting the agent to have the proper functionality. She defines the fundamental problem of agenthood as *action-selection*:

> Given an autonomous agent which has a number of general goals and which is faced with a particular situation at a specific moment in time. How can this agent select an action such that global rational behavior results? [Maes, 1989a]

Appropriately enough, her architecture, the Agent Network Architecture (ANA), is often nicknamed “Do the Right Thing.”

With this outlook, Maes still uses a behavior-based approach, but takes a different point of view on the question of how behaviors should be integrated. It is very unlikely that a designer will be able to foresee all possible combinations of events in the environment so that the agent will always take the right action. Instead, she wants to let her agents do some reasoning to figure out the best action to take, though she does not want to return to a system where reasoning dominates over action in the environment. In order to do this, she has developed a sophisticated action arbitration mechanism to let the agent quickly and mostly correctly decide which action it should take.

Maes divides her agents into “competence modules,” which basically correspond to behaviors for Brooks [Maes, 1990c]. A competence module is capable of taking some kind of action in the world, related to the tasks for which the agent is programmed. Competence modules are grouped according to how they relate to the overall goals of the agent. Competence modules basically act on their own, but they allow for low-bandwidth communication to decide which module should be active. All competence modules are always active, but they are only allowed to actually do something if they are activated using a spreading activation system.

Humanists, Maes’s technique of spreading activation is related to the hormonal system described on p. 32. It is loosely based on analysis of neural networks.

Specifically, modules are connected to each other according to the logic of their organization for a task. To put it a little too simply, modules have positive links with other modules that make them possible; and they have negative links with other modules that make them impossible. To start out, modules get “energy” if they are possible in the world, or if they are desired goals. Modules then spread energy over the positive links, and block energy over the negative links. The result is that, on average, the module that is most likely to help achieve the most important goal is chosen. [3]

Agent Network Architecture Agent Design Strategy

1. Choose a set of goals for the agent in its environment.
2. Identify tasks that will allow the agent to achieve each of the goals.
3. Break each task into its component actions.
4. Determine the preconditions and effects of each action
5. Determine how actions affect each other: which actions make other actions possible, which actions undo the work of other actions
6. Make links between actions according to how they affect each other

[3]This explanation is of necessity extremely simplistic. I apologize and refer interested readers to [Maes, 1989b].

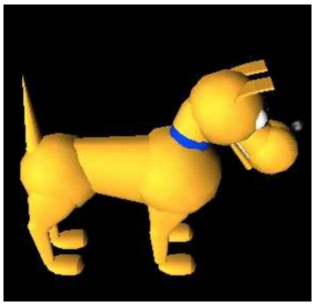
Hamsterdam

Bruce Blumberg builds on ANA by taking very seriously the notion of agent as animal [Blumberg, 1994]. Ethology — the study of animal behavior — has been an at least vague source of inspiration for many alternative AI researchers, but Blumberg integrates ethological principles into agent architecture to a new degree. His system, Hamsterdam, can be seen as a hybrid of the Maes goal-achieving optimality approach with ethological principles, based on the question: how can a creature, whether biological or artificial, decide, at each point in time, what is best for it to do?

Blumberg extensively adapts concepts from ethology in order to be able to build artificial creatures that share some of the properties ethology has identified as belonging to living creatures. For Blumberg, then, the 'units' of his agents are behaviors, as understood by ethologists. This means black boxes like "walk" or "sleep," with only simple interaction between them. Behaviors are related to drives or needs (hunger, fatigue) which they can fulfill. Behaviors are hierarchically organized into "behavior groups," which represent alternative ways to fulfill the same drive.

"[W]e wish to build interactive 'life-like' creatures such as virtual dogs, beavers, monkeys, and perhaps one day characters which are caricatures of humans. The fundamental problem for these creatures (as well as for their real world counterparts) is to 'decide' what among its repertoire of possible behaviors it should perform at any given instant." ([Blumberg, 1996], 29)

Blumberg's technique of combining behaviors is based on action-selection. The agent continuously redecides its actions, so that at any point in time the creature is engaging in the 'best' behavior (where 'best' is a combination of environmental appropriateness with factors such as maintaining a persistent focus of attention). Behaviors compete for control of the body, using a 'winner-take-all' scheme that works as follows: Behaviors constantly monitor the environment for conditions under which they might be appropriate. When they are triggered, they calculate a value that represents their appropriateness. Roughly speaking, the behavior with the highest value is allowed to take an action; 'losing' behaviors may suggest actions which the 'winning' behavior may or may not also take (for more details see [Blumberg, 1996]).


Blumberg's Silas

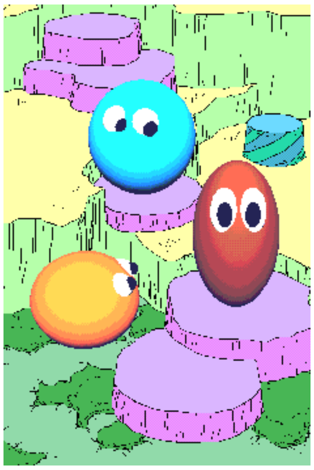
Hamsterdam Agent Design Strategy

1. Choose a creature in the world or a character to model
2. Decide on the needs and drives of the creature
3. Decide what behaviors the creature has, and how they fulfill the chosen needs and drives
4. Cluster related behaviors together into groups according to how they contribute towards the agent's actions
5. Manipulate the variables used to select behaviors in each group to get appropriate behavior under different circumstances
6. Manipulate the variables used to select behaviors between groups to get appropriate overall behavior in different circumstances

Hap

Loyall and Bates's Hap [Loyall and Bates, 1991] [Loyall, 1997a], the system on which this thesis work is based, is similar in many respects to Hamsterdam and a number of other reactive architectures. It is, however, the first such agent architecture to be focused on agents which are characters, rather than agents as animals or tools. Essential to the Hap understanding of agents is that an artist is attempting to express their vision of a particular character through the constructed agent. Conceptually, the Hap agent, while generally not amazingly intelligent, is "believable" as a character. This means it conveys a strong, author-chosen personality while avoiding doing anything so wrong that its audience is jarred out of belief in the agent as a living being.

" 'Believable' is used here in the sense of believable characters in the arts, meaning that a viewer or user can suspend their disbelief and feel that the character or agent is real. This does not mean that the agent must be realistic. In fact, the best path to believability almost always involves careful, artistically inspired abstraction, retaining only those aspects of the agent that are essential to express its personality and its role in the work of which it is a part." [Loyall, 1997b]


The Woggles

For Hap, it is not so important that the creature do the right thing with respect to fulfilling goals and drives in the environment. Rather, it is important that the agent be able to express its personality clearly down to the details of its behavior. At the same time, the agent must clearly react to what happens around it, appear to engage in goal-oriented behavior, be aware of what other characters and human interactors are doing, and in general not do anything that breaks the audience's suspension of disbelief. This means that the Hap architecture needs to combine the reactivity and environment-centeredness of other alternative AI architectures with a greater emphasis on author control of the details of behavior, rather than having behaviors be more or less generic, or having the details of the behaviors gradually emerge from what the agent decides to do.

The Hap architecture splits agents into goals and behaviors. Goals are simply names that represent to authors what the agent is doing (e.g. "dance").[4] Behaviors are intended as methods for doing goals, and they consist of author-written collections of physical actions (e.g. "jump") and other goals. Behaviors are made reactive by annotating them with environmental conditions under which they are or aren't appropriate to do; a behavior that is running will terminate itself when and if it becomes inappropriate.

When behaviors run, they can simultaneously start multiple goals. After some time, then, an agent may be pursuing quite a few goals simultaneously. Interaction between goals is handled by a priority mechanism, in which goals of high priority will be chosen over goals of lower priority. In addition, the author can mark particular combinations of goals as conflicting, so they can never happen at the same time. Additional details on Hap can be found in [Loyall, 1997a].

Hap Agent Design Strategy

1. Design a character to be implemented, including typical behavior and personality
2. Choose a set of high-level goals the character will engage in
3. For each goal, write a set of behaviors that instantiate that goal in different situations in a way appropriate to the character's personality
4. Each behavior may introduce new goals, so continue step 3 until all goals have behaviors
5. Add annotations to goals that conflict with each other

Summary: Alternative AI Agent-Building

Each of the listed architectures adds something important to the mix that is alternative AI. For the sake of the argument here, the following aspects are most important:

[4] Note this is different from the definition of 'Goals' given earlier.

- From Brooks comes the concept that agents should be divided into behaviors, each of which can run independently. Behaviors are basically independent, though there may be some low-bandwidth communication between them.
- Chapman and Agre introduce the idea that if an agent is to be situated responsively in an environment, it should redecide its behavior on every time step. By continuously redeciding behavior, the agent immediately responds to changing environmental conditions. These rapid alterations in behavior lead to the generativity of their architecture.
- Maes makes the critique that, for reasonably complex agents, decisions about how to arbitrate between behaviors cannot be made ahead of time. Therefore, agents will need to be able do some reasoning on their own. Maes introduces and Blumberg refines the concept of action-selection, i.e. that at every time step the agent should choose an action that is 'best' according to its goals or drives.
- Loyall and Bates add the concept that an agent should be written with an eye to how it affects its audience.

These architectures have disparate views of what an agent is, taken from different backgrounds: biology, ethnomethodology, engineering, ethology, and character animation. At the same time, a generally shared picture of agent construction emerges:

Note these are not the only commonalities between these architectures, or the only characteristics defining alternative AI. These are simply the ones most pertinent at this stage of the argument.

- Agents are seen as situated in an environment. Therefore, an agent's 'parts' are behaviors, which each may result in visible action in the world. Each behavior is firmly anchored to perception of the environment (when am I appropriate?) and to action upon the environment (what should I do?).
- Behaviors run relatively independently of one another. Each behavior does its own sensing, world modeling (where necessary), and makes its own decisions about appropriate action. Behavioral coordination and communication is minimal. All behaviors are running all the time, or at least when they are possibly appropriate. An agent may or may not actually simultaneously take actions caused by multiple behaviors.
- Conflicts between behaviors are handled with respect to what is most appropriate under given environmental conditions, and, for some architectures, with respect to what is most appropriate given current goals, emotions, drives, and / or recent actions.
- In order to remain reactive, agents continuously redecide their behavior in light of changes in the environment (as well as changes to their internal state).

Schizophrenia as a technical problem

One of the fundamental complaints alternative AI makes about classical AI is that it focuses on the functional components of intelligence. These components are generally hard to integrate into a complete agent. Their underintegration can manifest itself, for example, in various kinds of inconsistency between the different functions, such as not being able to use knowledge for one function that is available for another. So the agent may speak a word it cannot understand or visibly register aspects of the world that do not affect its subsequent behavior.

In contrast to this functional decomposition, alternative AI designs behaviors, each of which integrate the functionalities it needs to operate. This does not solve all problems, however, since the various behaviors also need to be integrated. Brooks, for example, has stated that one of the challenges of the field is to find a way to build an agent that can integrate many behaviors, where he defines many to be more than a dozen [Brooks, 1990]. In complex agents that exhibit many behaviors, those behaviors are extremely difficult and tedious to integrate completely, with the result being that they often remain only loosely integrated.

The reason for this difficulty can be traced to a fundamental tenet of the behavior-based approach. The design choice in behavior-based AI is to build behaviors independently, and have them use minimal communication and coordination. This black-boxing approach has the advantage of simplifying agent design, since each behavior can be designed and built separately. It can also give you a complete, though limited, agent sooner, since each behavior is in effect a complete agent. Nevertheless, the black-boxing approach raises the question of how the different behaviors of the agent can be made to work together properly. The next section gives a concrete example of these problems; this will put us in position to formally define the difficulties of integration for alternative agents.

Case Study: Integrating the Woggles

In 1992, a group of 13 researchers, including Oz Project members, built "The Edge of Intention" [Loyall and Bates, 1993] (Figure 2.6, a system containing small, social, emotional agents called "Woggles" that interact with each other and with the user. We used the Hap architecture to build these agents.

Following the Hap design strategy, we first built a set of high-level behaviors such as sleeping, dancing, playing follow-the-leader, moping, and fighting. Each of these behaviors was reasonably straightforward to implement, including substantial variation, emotional expression, and social interaction. Watching the agents engage in any of these behaviors was a pleasure.

Then came the fatal moment when the behaviors were to be combined into the complete agent. This was a nightmare. Just combining the behaviors in the straightforward way led to all kinds of problems:

- Agents would try to engage in two behaviors simultaneously that did not make sense (e.g. , 'fight' and 'sleep' — we optimistically called the result "emergent nightmares").

- Agents would switch from one behavior to another with their body in an unusual state. For example, an agent startled out of sleeping might engage in several subsequent behaviors with its eyes shut.

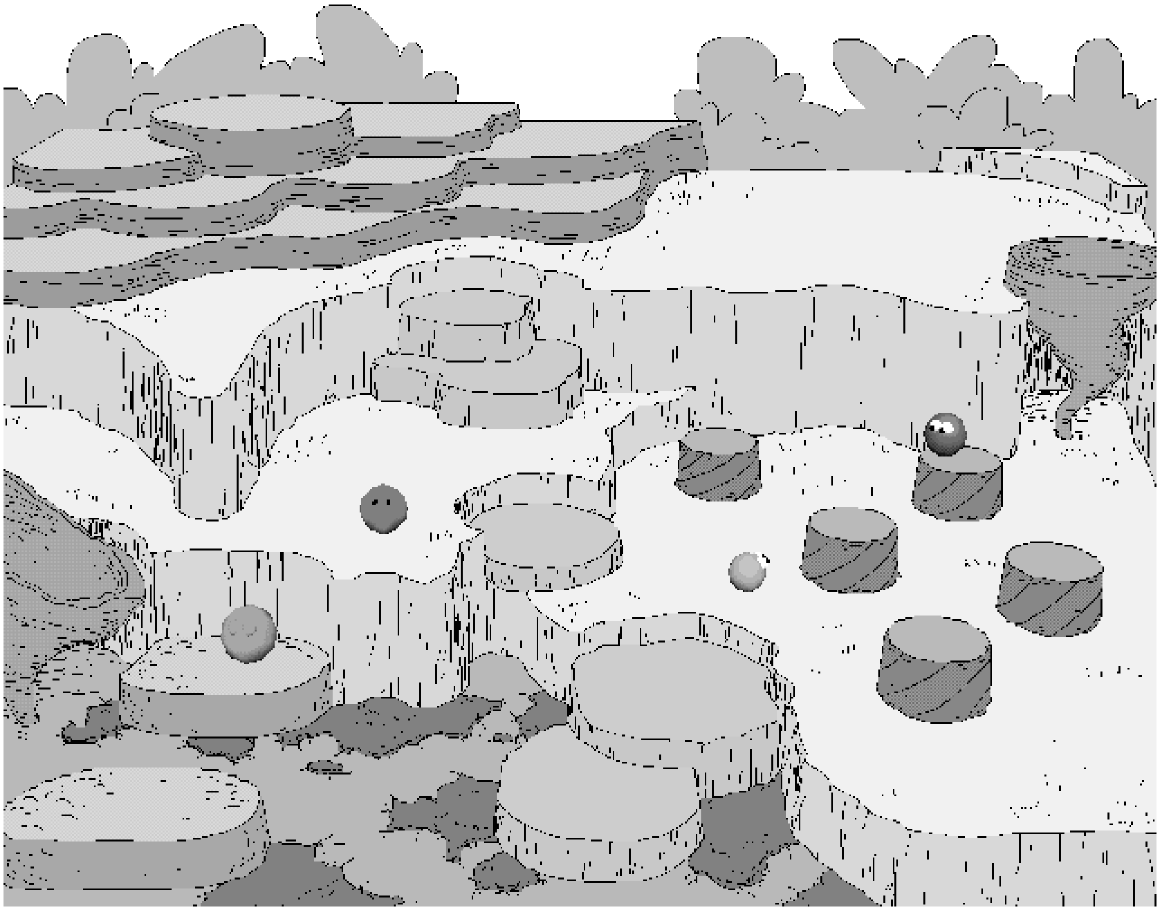


FIGURE 2.6: The Edge of Intention

- Agents would rapidly switch from behavior to behavior, never settling into one long enough to make the resultant activity comprehensible. Alternatively, agents would refuse to switch from one behavior to another in situations where they really should change, making it seem that the agent was clueless about what it was doing.
- Agents would get stuck in 'loops' where they kept switching back and forth between two behaviors, never being able to settle down into one until something in the environment drastically changed.

Under the pressure of deadlines, we added an ad hoc system to handle interbehavioral coordination: agents could only engage in one high-level behavior at a time; when switching from behavior to behavior, we reset the most crucial aspects of the body (open the eyes, stop trembling, stand up straight, etc.); express personality by varying the probability that you would engage in a particular high-level behavior under circumstances where it is appropriate. This clearly improved matters, but it did not fundamentally solve any of the problems, and they still regularly reared their ugly heads during runs of the system.[5] While individual behaviors

[5]Loyall believes that many of these problems were rooted in a bug in the way in which

were easy to write, the interactions between behaviors — particularly as manifest in the agent's apparent activity — were difficult to control and manage properly. These problems are not unique to Hap.

Schizophrenia defined

Alternative AI, while clearly having some impressive results, has not solved a fundamental difficulty of classical AI, i.e. its inability to integrate the parts of the agent into a coherent and coordinated whole. Generally, the agent's behaviors are too crystallized; the boundaries between the agent's behaviors are too sharp. Unlike biological agents (or characters in a film, for that matter), one can see the boundaries between the agent's behaviors.

In particular, alternative AI agents generally have a set of black-boxed behaviors. Following the action-selection paradigm, agents continuously redecide which behavior is most appropriate. As a consequence, they tend to jump around from behavior to behavior according to which one is currently the best.[6] What this means is that the overall character of behavior of the agent ends up being somewhat deficient; generally speaking, its behavior consists of short dalliances in individual, shallow high-level behaviors with abrupt changes between behaviors. It is this overall defective nature of agent behavior, caused by under-integration of behavioral units, that I term *schizophrenia*.

Because all behavior-based systems do not integrate behaviors in the same way, they also do not exhibit schizophrenia in the same way. Some of the difficulties with Hap are noted above. Each of the other architectures has its own style of schizophrenia, which is best observed visually or through the experience of programming, but can sometimes be gleaned from research reports.

- Brooks's experience seems to parallel ours with the woggles. Adding new low-level behaviors to his robots is straightforward using the subsumption technique. However, Brooks does not even try to integrate many high-level behaviors; he states up front that it is not currently possible. Getting coherent overall behavior is an open question: "A humanoid robot has many different subsystems, and many different low level reflexes and behavioral patterns. How all these should be orchestrated, especially without a centralized controller, into some sort of coherent behavior will become a central problem" ([Brooks, 1997], 297).

- Pengi jumps from action to action according to whatever seems most appropriate from moment to moment. As a consequence, Pengi mixes its behaviors together in ways that may or may not result in activity that seems to make sense. As Agre and Chapman charmingly put it, "Pengi regularly... combines its repertoire of activities in useful ways we didn't anticipate. (It also regularly does silly things in situations for which we haven't yet wired it)" ([Agre and Chapman, 1990], 23).

conflicts between goals were handled [Loyall,]. Subsequent experience by other designers with a debugged version of the system ([Neal Reilly, 1996], Chapter 7 of this document) suggests that while this may have been part of the story, substantial problems remain.

[6] A similar observation is made by Steels [Steels, 1994].

- Like Pengi, ANA's schizophrenia manifests itself in jumps from action to action. Nevertheless, the silliness of the results is mitigated somewhat by the fact that the system itself is doing some reasoning about what is appropriate to do. It is a little difficult to judge it accurately, though, without being able to see the dynamics of the system in action, preferably on a set of complex high-level behaviors.

  Some provisional conclusions may be drawn from Bradley Rhodes's PHISH-Nets [Rhodes, 1996], which built on ANA, and was used to implement several characters. These characters displayed a kind of schizophrenia where they could reason extensively about conditions in the environment, but then moved abruptly and rather woodenly from atomic behavior to atomic behavior. While this may be more an indicator of the limits of a master's thesis system than an inherent characteristic of ANA, it seems likely that, if used to drive a graphically represented agent, ANA would have the tendency, like Hap, to switch rapidly from behavior to behavior, and to get stuck in behavioral loops.

- Silas, the dog built using Hamsterdam, is like Pengi in jumping from behavior to behavior. Unlike Pengi, Silas's individual behaviors are well-integrated, so it is fairly clear which behavior Silas is engaging in. Unfortunately, this increase in behavioral coherency also increases Silas's apparent schizophrenia, since it leaps from behavior to behavior, in a way that is clear and can be abrupt and jarring. Often, there is no clear thread connecting the behaviors, resulting in an appearance of either behavioral randomness or pure stimulus-response.

While schizophrenia manifests itself in different ways, it can generally be understood as a manifestation of the limit point of behavior integration. Programmers can create robust, subtle, effective, and expressive behaviors, but the agent's overall behavior tends to gradually fall apart as more and more behaviors are combined. For small numbers of behaviors, this disintegration can be managed by the programmer, but as more and more behaviors are combined their interactions become so complex that they become at least time-consuming and at worst impossible to manage. Schizophrenia is the symptomatology by which behavioral underintegration can be directly observed in the agent. It manifests itself in at least two ways that make the resulting system hard to understand: (1) switching abruptly and mechanically from high-level behavior to high-level behavior; (2) mixing actions from different behaviors together in an incoherent jumble .

Why schizophrenia?

At this point, you may be wondering to yourself why on earth I am using the psychiatric term 'schizophrenia' for this technical difficulty. If so, good for you! Schizophrenia is a complex term, loaded with a history of contradictory uses and abuses in a variety of fields, and so full of 90's cultural theory cachet that observers may wonder if it really still means anything at all.

It is with some trepidation, then, that I introduce this term now in a technical context. I believe that the reasons for its use in this case are so compelling that they outweigh the dangers of adding to the obfuscation that already exists around this term. In particular, many uses of the term schizophrenia bear deep relations with the problem of behavioral underintegration in alternative AI. Giving these usages the same name allows for the development of their metaphorical connections, making it potentially illuminating to look at all these versions of schizophrenia simultaneously. In this respect, focusing on this technical problem as a variation on schizophrenia may actually increase understanding of schizophrenia, rather than further diluting the term.

While an extensive examination of schizophrenia will have to wait until Chapter 3 and Intermezzo I, I will here explain how schizophrenia has historically been used in psychiatry and cultural theory, and clarify how it relates to current problems in AI. The key point will be the multiple uses of schizophrenia as a metaphorical concept, and how they each put the difficulties of alternative AI in a new light.

1.Schizophrenia as incoherence

The notion of schizophrenia as a psychiatric term is generally seen as originating with Kraepelin, who unified a variety of disorders under the name *dementia praecox* in 1898. This name was intended to refer to the fact that these disorders all seemed to be related to a gradual mental deterioration that began when the patient was young. In 1911, Bleuler renamed this heterogeneous group of disorders *schizophrenia* "because he thought the disorder was characterized primarily by disorganization of thought processes, a lack of coherence between thought and emotion, and an inward orientation away from reality. The 'splitting' thus does not imply multiple personalities but a splitting within the intellect and between the intellect and emotion" ([Coleman *et al.*, 1984], 344). The usages of the term schizophrenia have tended to cluster around the description which Bleuler gave, specifically emphasizing an internal incoherence and disorganization. The incoherence we see in alternative AI agents, then, can be put in a broader light: it corresponds at a high level with some conceptions of schizophrenia from psychiatry. This will be the most basic, and most inaccurate, usage of schizophrenia here.

2.Schizophrenia as a meta-level incoherence

Schizophrenia has never been a straightforward, easily identifiable syndrome. The heterogeneity of the disorders and symptomatology to which the term schizophrenia can be applied has led to a substantial amount of diagnostic creep in this "most baffling" ([Coleman *et al.*, 1984], 345) of psychiatric disorders, including substantial variation between geographical regions and over time. The Diagnostic and Statistical Manual of Mental Disorders (DSM) [American Psychiatric Association, 1980], the official repository of definitions of mental illness, has reflected these variations.

> [T]he criteria are a curious mixture of an older set of concepts originally proposed by Bleuler (1911, 1950) and a newer set, chiefly those of Schneider (1959), which appear to have only an obscure and unspecified relationship to each other. In consequence, we cannot be sure that persons on whom much of our research knowledge depends — those

> who were diagnosed as schizophrenics under, say, DSM-II — can be grouped with persons described as schizophrenic under DSM-III. In one study peripherally concerned with this dilemma, a group of 68 DSM-II-defined schizophrenic patients was reduced to 35 when DSM-III criteria were applied, a reduction of 51 percent! ([Coleman *et al.*, 1984], 353)

Even using one particular criterion, the concept of schizophrenia is hard to pin down, with proliferating subcategories, symptoms, and relations, rather than a set of properties with a common core. "There is, in fact, no constant, single, universally accepted 'sign' of the presence of schizophrenia;" ([Coleman *et al.*, 1984], 354) this psychological textbook concludes that the only common feature is behavior that is bizarre and unintelligible ([Coleman *et al.*, 1984], 353).

What's interesting, then, is that schizophrenia refers to a kind of incoherence, but is itself incoherent as a concept. It is notoriously difficult to pin schizophrenia down as a particular thing, a fact which reflects itself in the multiple metaphorical uses I list here. It is equally difficult to classify people accurately and repeatedly as schizophrenic. Schizophrenia in psychiatry, then, can also be understood as a meta-level problem: the difficulty of understanding and classifying people within a rational system. Schizophrenia in this sense represents the limits of our ability to categorize people.

Categorization enters into alternative AI as well: the first step of designing an agent requires us to divide the agent's overall, perhaps ineffable behavior and personality into a set of relatively clearly-defined behaviors. Schizophrenia as meta-level incoherence suggests that this step is fraught with danger, since there may be limits to our ability to understand and categorize behavior and those limits may manifest themselves in incoherence at the level of synthesis. The concrete implications of this for alternative agents will become more apparent in the analysis of agent construction later in this chapter.

3.Schizophrenia as a theory of consciousness

As noted in Chapter 1, schizophrenia for cultural theorists refers to a particular way of thinking about what it means to be human in contemporary Western society. This usage came about in response to perceived difficulties with the rational model of subjectivity. This is because the rational model no longer works when we talk about people who have traditionally been marginalized. If the rational is the definition of what it is to be human, it is equally true that disenfranchised people, such as women and blacks, have often been classified as nonrational and hence as unworthy of the status of full humans. For example, when we deal with the mentally ill, we are dealing with people who by definition are nonrational [Foucault, 1973].

The use of the term 'schizophrenia' to describe a kind of subjectivity that could apply to everyone — not just the mentally ill — is inspired by the antipsychiatric movement of the 1960's. The antipsychiatrists seek to include those with mental illnesses in the category of the 'human' by describing their mental processes as simply more extreme versions of processes that take place in everyone's mind, rather than as the fundamentally different (nonrational) way of thinking the rational model has to assume. In short, a schizophrenic model of consciousness takes into account the experiences of (for instance) the mentally ill in order to create a more inclusive model of human experience.

"[W]e believe that abstract reasoning is *perforated*: it is not a coherent module that systematically accounts for all or even a class of mental phenomena. It is not a general-purpose reasoning machine, as it appears to be, but only a patchwork of special cases." ([Chapman and Agre, 1986] 415)

"The existential concern that animates our entire discussion in this book results from the tangible demonstration within cognitive science that the self or cognizing subject is fundamentally fragmented, divided, or nonunified." ([Varela *et al.*, 1991], xvii)

This antipsychiatric belief that normal human existence is fundamentally irrational and incoherent, with a thin veneer of apparent rationality and cohesion mostly supplied by self-delusion, is also common in many alternative AI writings. Alternativists Francisco Varela, Evan Thompson, and Eleanor Rosch relate the disunified self of enactive cognitive science to that understood by Buddhism, stating that both meditation and cognitive science uncover the nonunity of consciousness. Brooks discusses in detail and on scientific grounds why our introspected view of consciousness as unified is fundamentally erroneous ([Brooks, 1995]). In general, the belief that agents can or should consist of separate behaviors with minimal interconnection easily leads to the conclusion that unity, rationality, and coherency are an illusion, or, at best, an emergent property of a fundamentally schizophrenic system.

At heart, antipsychiatrists do not believe that schizophrenics are fundamentally different from other people. As a consequence, they consider 'schizophrenic' as a label to be inaccurately applied to a single person. Rather, antipsychiatrists understand schizophrenia as a social or interpersonal problem; they may claim, for example, that schizophrenics are responding sanely to an insane environment. Fundamentally, they see schizophrenia as an interaction between a person and his or her surroundings. While more recent studies suggest that schizophrenia is not purely or perhaps even largely environmental, the notion that mental illness can be profitably understood by situating a patient in the context of their environment has remained current [Minuchin *et al.*, 1978].

This belief in schizophrenic consciousness as situated in an environment parallels alternative AI's insistence that intelligence can only be understood in terms of environmental interaction. Both antipsychiatrists and alternative AI researchers believe that behavior does not exist in a vacuum. According to this viewpoint, behavior can only be fairly evaluated by understanding it as an interaction between an individual and his or her environment.

4.Schizophrenia as a consequence of a particular kind of decomposition of subjectivity

For cultural theorists, 'schizophrenia' is considered to be both a general way of thinking of people in the 20th century, and a particular and not necessarily positive way of being that is a result of the largely technological and industrialized world in which we live. Schizophrenia is here understood to be a result of living under a system where people are engaged only in terms of one part of their personality; over time, they lose their cohesion as different parts of the personality become autonomous and are no longer coordinated with one another. The paradigmatic example of this kind of schizophrenia is the worker on the assembly line, who may undergo exquisite psychic torture as he or she performs repetitive, mindless motions all day [Doray, 1988].

"With the modern 'psychological' analysis of the work-process (in Taylorism) this rational mechanisation extends right into the worker's 'soul': even his psychological attributes are separated from his total personality and placed in opposition to it so as to facilitate their integration into specialized rational systems and their reduction to statistically viable concepts"([Lukács, 1971], 88)

Schizophrenia in this sense is the limit point of rationalization as it is applied to human consciousness. It is understood as a kind of disintegration that comes about as all qualitative aspects of humanity are eliminated, to be replaced by quantitative, autonomous, and individually rationalized units. With this analysis of schizophrenia as a result of decomposition, we have come full circle: this style of schizophrenia corresponds directly to the technical difficulties alternative AI practitioners face in getting the parts of their autonomous agents to act coherently. Alternative AI practitioners, too, split the 'souls' of their agents into autonomous, quantitative units; their agents suffer from the same kind of 'schizophrenia' cultural theorists have identified in modern humans. The broader implications of this cultural theory understanding of schizophrenia for alternative AI practice will be studied in greater depth in Chapter 3.

Summary of schizophrenia as metaphor

Each of the metaphors of schizophrenia forms a strand which connects formerly disparate intellectual practices. The strands are summarized in the table below. The advantage of using the term 'schizophrenia' is that by studying these strands together, each area has the chance to shed light on the other. At the same time, it is important to note that the usage of schizophrenia in this thesis is not intended to be final. Schizophrenia is not only a metaphor, but also a serious syndrome that affects many people's daily lives. Its usage here is not meant to make light of their suffering or to suggest treatment options.

From domain	Schizophrenia as...	Corresponds to...
Psychiatry	incoherence	incoherence of behaviors
Philosophy of science	meta-level incoherence	problems in understanding agents
Anti-psychiatry	theory of consciousness	concept of agent
Cultural theory	related to psychological decomposition	behavioral decomposition

Does schizophrenia matter?

Many postmodern theorists have achieved a comfortable notoriety by antagonizing more traditional theorists with their celebration of the virtues of schizophrenia. Simply put, schizophrenia represents for them a liberation from the constraints of behaving as a rational, repressed, neurotic individual. Similarly, many believers in alternative AI celebrate the schizophrenia inherent in their agents. By getting away from a central, hierarchical organization, these scientists feel that they are getting away from many of the flaws of classical AI, and the resultant schizophrenia in their agents becomes a proud marker of their rejection of classical ideas of agenthood.

On the surface, you may find this attitude, if not correct, at least reasonable. Abrupt switching between homogeneous behaviors does not seem such a terrible flaw in the overall scheme of things. Here I will argue that, in fact, schizophrenia can be a fundamental problem, depending on the use to which complex autonomous agents will be put.

The problems schizophrenia raises depend on the use to which you would like to put your autonomous agent. Clearly, for some uses, schizophrenia does not matter at all. If a vacuum-cleaning robot jumps from its vacuuming to its wandering-about-the-house behaviors, this probably does not degrade its vacuuming duties.

Suppose, though, that you want to use an agent as a believable agent, i.e. as a character to which a human is supposed to be able to relate. Believable agents are supposed to allow for a suspension of disbelief, and when they remain too rigidly in one behavior, or switch abruptly between behaviors, they seem unnatural. In addition, agents with a small set of rather shallow behaviors are not so engaging; the user quickly learns to identify the major behaviors the agents can engage in, and then interaction reduces to getting the agent to do one of its 'tricks.' Making agents less schizophrenic means, for believable agents, that the set of behaviors with which the agent is programmed are not so transparently obvious to the user; that the 'parts' from which the agent is built, its behavioral units, blend into a whole personality which invites exploration and discovery without immediately exhausting it.

You may also want to use your agent as a scientific model of a living creature. When we are attempting to build a model that behaves in a similar way to living agents, schizophrenia is something of a problem. After extensive, if not entirely scientific, observation of living agents in the world [Sengers,], I have found it impossible to exhaustively identify the set of high-level behaviors in which the agent engages, and I only very rarely notice abrupt switching between clearly-defined high-level behaviors.

Why I chose not to do a scientific study will become clearer in Chapter 6.

What I *have* noticed is that the very search for high-level behaviors tends to consist of watching a conglomeration of somewhat undifferentiated activity and attempting to come up with plausible explanations about what the agent is doing. What this implies is that the whole notion of 'high-level behavior' is a convenient explanatory mode for identifying gross animal behavior, but that it does not have a necessary detailed correspondence to what the agent is 'actually' doing. The agent may be engaging in a lot of low-level behavior that does not correspond to any high-level behavior, or it may be engaged in some ineffable behavior to which we can simply attach various explanations. When we build scientific models that allow for easy identification of the gross behaviors in which the animal engages, those models are inaccurate in that they display features which living agents do not display, features which are purely a result of the way we built our model. Schizophrenia, not being an attribute of animals in the way we have defined it here, is therefore a problem for scientific agents as well as believable ones.

You may not care about scientific correctness, but simply want to use your agent as a tool. In this case, coherence in the agent is not a value to be achieved for its own sake; it does not bother me, for example, that my text editor switches abruptly from its "writing" to its "printing" modes. However, there are many times when it is not enough for the agent's actions to achieve the user's goal; the user must also be able to understand *why* the agent does what it does. If, for example, a person is teleoperating a semi-autonomous robot, it may be very important that the person can quickly and easily understand what the agent is doing by watching it.[7] If the agent is changing abruptly from behavior to behavior, or switching behaviors so rapidly that the user cannot figure out what the robot is doing, teleoperating it will become much more difficult. Schizophrenia matters for agents-as-tools, because these tools are complex and are often

[7] I am indebted to Red Whittaker for this example.

used by people who need to be able to understand what they are doing.

Finally, you may simply be an AI dreamer who wants to be able to build creatures that are as engaging and interesting as biological beings. If this is the case, you should find schizophrenia very upsetting. Schizophrenic agents do stupid things; they look bad; they reveal the very fact of their mechanicity at every turn. At least a minimal level of coherence is an essential part of what we (albeit perhaps incorrectly) attribute to intentional agents. We will look at this phenomenon in more detail in Chapter 6.

If you are a humanist who does not care about AI, you can sit back and smile politely. Don't worry, your time will come in Chapter 8.

The root causes of schizophrenia

While this formulation of schizophrenia is new, the problem of behavioral coherence has long been recognized in alternative AI. At its most basic, the problem of integrating multiple behaviors per se is foundational. More specifically, building agents that are coherent - that appear to behave consistently across goals and behaviors, not as a bundle of parts — is an explicit goal for many researchers. Brooks ([Brooks, 1994]), Blumberg, and Loyall, for example, all mention this kind of apparent behavioral coherence as a goal of their work.

> "Even though many behaviors may be active at once, or are being actively switched on or off, the robot should still appear to an observer to have coherence of action and goals. It should not be rapidly switching between inconsistent behaviors, nor should two behaviors be active simultaneously, if they interfere with each other to the point that neither operates successfully" ([Brooks, 1991a], 22).

These researchers have put a lot of work into trying to understand how to design and build coherent agents. Loyall has, for example, developed agent design strategies and architectural support for mixing multiple activities in pursuit of a goal, so that an agent does not, for example, freeze in place while it is trying to decide what to say. Blumberg has also put substantial effort into addressing coherence. His system addresses the problems of rapid switching and multiple conflicting behaviors, and he has some novel techniques for combining simultaneous behaviors. However, the problem of abrupt behavior switching remains, and is, if anything, more clear in his systems than in the others. Fundamentally, these solutions, while chipping away at particular symptoms of schizophrenia, do not address the fundamental problem that behaviors are designed separately, and that the boundaries between them become clear in the activity of the agent.

Given this recurrent interest, the inevitable conclusion must be that the problem of schizophrenia has not remained unsolved due to a lack of interest, effort, or talent. Why is this problem so hard to solve? To put it at its simplest, behavior-based AI runs into the same problems classical AI has — if you divide your agent into parts, it is natural to have problems integrating those parts back together again. But the problem is deeper than finding some ad hoc solution to hook up the disparate parts of any particular architecture. The claim I will make here is that *schizophrenia has not been solvable because it is an inevitable result of the current agent design process*. In order to understand this, we will need to take a closer look at how we construct agents.

Agents as atomized

Unlike biological agents, artificial agents begin life as a concept in their designer's head. At this stage, an agent is an idea — a living agent to be copied, a dreamed-up character, a potential solution to a problem — which is analyzed in order to yield the constituent parts that become the eventual agent's behaviors.

"There are many possible approaches to building an autonomous intelligent system. As with most engineering problems they all start by decomposing the problem into pieces, solving the subproblems for each piece, and then composing the solutions." ([Brooks, 1986b],6)

For example, imagine that you want to build an artificial dog. To do this in the behavior-based manner, you will first need to decide what the basic behaviors of the dog are. You generally do this by looking at the dynamics of the agents' activity, and trying to recognize what the parts of that activity are. Thinking about or watching a dog, you might come up with some typical behaviors: eating, playing fetch, sleeping, etc. After you've come up with the behaviors, you connect them using your architecture's default behavioral organization mechanism. In the end, you might end up with something like Figure 2.7.

Behavior-based agent design works by breaking the dynamics of the imagined or observed interactions into parts by parsing the dynamics of the agent's behavior for meaningful subunits. That is, the behavioral units chosen for the agent are a result of an interpretation of the imagined or observed agent's interactions with a user or environment. This interpretation is fundamentally symbolic; as in parsing, the agent's behavioral dynamics are divided into meaningful, somewhat independent units.

If you have a humanist background, you may recognize atomization as a form of reification, applied to objects of scientific study. This view of atomization will be explored in more detail in Chapter 3. Atomization is also similar to reductionism, the belief that objects are made up of the simple combination of simpler objects. Atomization is, however, not necessarily a statement about the way the world is organized; it can simply be a way of approaching phenomena in order to make them easier to understand.

This process of splitting-up I term *atomization*. Strictly speaking, atomization refers to the process of splitting something that is continuous and not strictly definable into reasonably well-defined, somewhat independent parts. The term *atomization* comes from neurology [Goldstein, 1995], where the *atomistic method* refers to a method of trying to divide the brain into small, localized pieces, each of which corresponds to exactly one behavior. The use of atomization in computer science, under such watchwords as modularity, decomposition, and divide-and-conquer techniques, has been more successful. These techniques form the core of programming methodology and are essential tools for making large systems that people can design and understand.

In fact, methodologies akin to atomization are not limited to computer science. The advantages of using atomization to understand complex systems are understood in many sciences. It has similarities, for example, with the digitization of analog signals, with dissection of organisms in anatomy, with the identification of species in population biology, with the classification of mental illness in psychiatry, and, in general, goes hand-in-hand with formalization and analysis. In all these cases, atomization is a way of getting a handle on a complex phenomenon, a way of taking something incoherent, undefined, and messy and getting some kind of fix on it.

It should be clear at this point that a fundamental tenet of behavior-based AI is behavioral atomization. Manifestos on behavior-based AI regularly cite behavioral decomposition into independent units with limited interaction as one of the defining characteristics of the movement:

> An agent is viewed as a collection of modules which each have their own specific competence. These modules operate autonomously and are solely responsible for the sens-

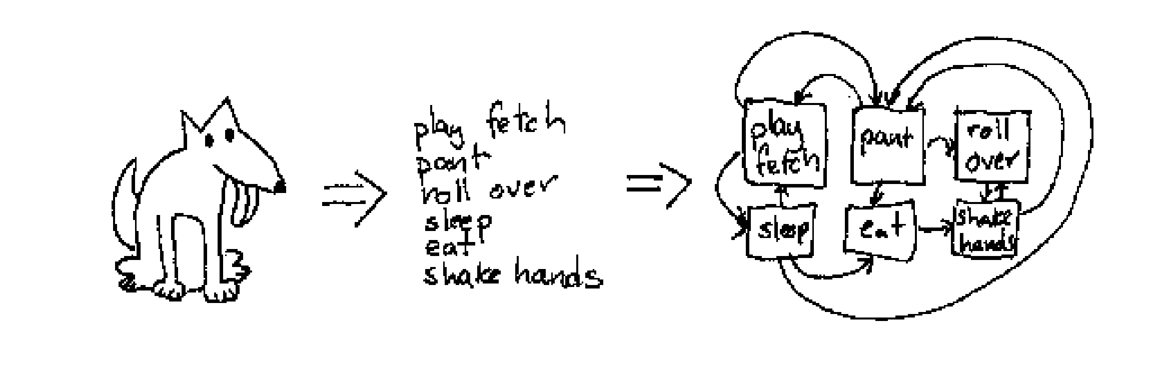


FIGURE 2.7: Design steps for an artificial dog

> ing, modelling, comptuation or reasoning, and motor control which is necessary to achieve their specific competence.... Communication among modules is reduced to a minimum and happens on an information-low level. ([Maes, 1990b],1)

Alternative AI researchers are generally allergic to the concept of central control; one generalization to which this not infrequently leads is that behaviors should be designed and built separately and refer to each other as little as possible.

Special properties of atomization

> "Real biological systems are not rational agents that take inputs, compute logically, and produce outputs. They are a mess of many mechanisms working in various ways, out of which emerges the behavior that we observe and rationalize." ([Brooks, 1995], 52)

In alternative AI, atomization is a process of breaking down a behavior into meaningful units closely akin to the process of parsing natural language. There is, however, a major difference between parsing natural language and understanding an agent's behavior; while in listening to native language speakers we can be reasonably certain that the stream being parsed truly contains symbols, it is unclear in what sense we can truly say an agent's physical presentation is a more or less linear stream of clear-cut behaviors. When observing living creatures, for example, one can certainly deduce a set of high-level behaviors [Benyus, 1992], but one would be hard-pressed to even identify every movement of an animal as being part of one of these behaviors, let alone understand all behavior purely as a succession of these well-defined, a priori behaviors. Given the mess that is the nervous system, it's hard to even imagine how such a neat, tidy behavioral presentation could ever happen.

Atomic behaviors, then, are not pre-given — they do not exist in the world per se. Rather, these atoms are an interpretation of agent activity, distilled into units which carry meaning for the observer. Atomization is a kind of explanation, a process of understanding that comes about as we try to bring order to our experience of the world. In this sense, atomic behaviors are not what the animal does, but our best explanation to ourselves of what the animal is doing. Atomization is a form of approximation, taking noisy, messy, real-world activity and distilling it into a more formal and clean representation.

This does not mean that atomization is arbitrary or useless. Atomization brings with it properties that are valuable; the representations it generates are essential to helping us understand and engineer behavior. Atomization is essential to science because it brings order to inchoate experience, giving you pieces out of which a formal system can be built. Atomization is, in fact, also endemic to the humanities (through theoretical categories such as race, gender, and class distinctions, literary periods, and genre), although the use of atoms there is somewhat different because of a greater ingrained skepticism for the absolutism of categories and because of a different conceptualization of how atoms relate to one another and to the phenomenon they model (we will build on this in Chapters 3 and 6). In both science and the humanities, atoms give you the basic units out of which meaningful understanding of the world can be constructed.

Atoms are not, however, simply transparent lenses through which the world is viewed. The atoms out of which we build our agents have their own special properties, which form the basis of our ability to use them to understand and build artificial agents:

- Atoms are **discrete**. Natural behaviors generally blend together, making it hard to define a clear moment when an animal changes, for example, between being asleep and being awake. Atomic behaviors partition these analog behavioral changes into clear states. There are no in-between states, no processes of transformation between one behavior and another, no moments when the action of the agent cannot be attributed to one of the labelled behaviors at all.
- Atoms are **meaningful units** of action. Atomic behaviors correspond to activities that make sense to the observer / designer. In this sense, behaviors are *symbolic*. They are conceptual chunks of the agent's activity.
- Atoms are **cleaner than real-world behavior**. Real-world behavior is messy, not always clearly definable or understandable. Atomic behaviors clean up this mess, allowing us to build systems that are understandable, programmable, controllable.

Atomization, then, is fundamentally *the reduction of an observed, analog stream of activity to discrete, meaningful, symbolic parts*.

Schizophrenia as a consequence of atomization

Because atoms have their own special properties, an agent that we build from atomized units is in important ways not equal to the thing it reproduces. An atomized agent consists of a symbolic representation of the original creature's actions. There is nothing wrong with this situation per se; it is a simple statement of the fact of the agent's construction. In order to understand the creature's actions, we create a symbolic representation of those actions; it is these symbols that form the basis for the engineered reproduction of the agent.

What happens when we build our agents from these symbolic units? The tendency is for one of two things to happen: either the behavior is completely incomprehensible to the user, or, to the extent that the behavior is comprehensible, the user can recognize the behaviors with which we programmed the agent. Since, generally speaking, we intend