for users to be able to recognize these units, this recognition is exactly what we wanted. And since the behaviors we chose are precisely what we found meaningful in the original creature's behaviors, it is not surprising when the interactor can recognize them in the agent as well. We are happy if we have succeeded, and the user knows it is hunting, playing fetch, etc.[8]

There is, however, a problem, in that all the actions of the agent are a result of these symbolic, hopefully recognizable behaviors. The agent's behavior, if understandable at all, becomes so clear-cut that it is amenable to a kind of parsing on the part of the observer that is unreasonable for living creatures. This is what causes them to seem unnatural. While with living creatures there is always some amount of 'noise' (the extent to which the atomizing approximation is only an approximation), artificial creatures are all high-level, potentially understandable, symbolic behavior. People quickly notice the categories into which the agent's activity is divided; they can see that, unlike biological agents, this creature is pure representation. This, then, is the source of schizophrenia in agents: the modularity of agent design into symbolically meaningful units means that the individual behaviors of an agent are too clear-cut. Agents jump from behavior to behavior in a jarring and often meaningless sequence.

At this point you may come to the conclusion that the way to solve this problem is not to have any explicit behaviors. This is in fact the solution used in architectures like Pengi and ANA. Agents built in these architectures do not exhibit any pre-planned behaviors per se, but rather mix together actions from different behaviors according to whatever seems logical at the moment. Interactors certainly will not recognize behaviors if there are no behaviors to recognize.

But this does not fundamentally solve the problem of schizophrenia. Agents do not engage in clear-cut behaviors, but mix together actions from different behaviors. Still, each action the agent chooses is from a particular, designed high-level behavior. While Pengi and ANA allow the agent to interleave actions — choosing actions alternately from different behaviors — they do not allow agents to engage in action that is not directly related to one of the designer's chosen high-level behaviors. This also means there is no mediation, averaging, or transformational processes between behaviors. The agent can only take action that is logical within the parameters of one of its behaviors.

In addition, in these architectures each action the agent takes is chosen because of its logic for a particular high-level behavior. Since actions are chosen for their logical role in separately designed behaviors, it is likely that they will make less sense in the agent's 'emergent' behaviors. In particular, strange behavior will result any time the logical structure of the two high-level behaviors is different. Since the philosophy of behavior-based AI is to design behaviors as separately as possible, this state of affairs is bound to happen regularly.[9]

[8] In addition, we are also sometimes happy if the user comes to think the agent is doing other intelligent things that, strictly speaking, we haven't programmed it to do. The phenomenology of projected behavioral identification would be an interesting subject for another thesis, to which I think Chapter 6 provides some initial clues.

[9] It is also not clear that the approach of behavior-less behavior construction scales well to multiple, complex behaviors. It may be that truly complex behavior of the kind required to run an articulated graphical or robotic agent with a wide range of activity requires structure like behaviors for the programmer to be able to keep track of what is going on.

Atomized behavior, then, if comprehensible at all, will display one or both of the following attributes, depending on how it is synthesized:

1. It may be synthesized in terms of symbolic behaviors. In this case, it will tend to jump abruptly from behavior to behavior. These behavioral switches will be apparent and jarring to the user, since users, like designers, will understand the agent's activity in terms of symbolic activities.

2. It may be synthesized at the level of actions, with actions chosen according to the logic of the separate symbolic behavior of which the designer sees that action as a part. In this case the actions will often be mixed in a way that violates the individual behaviors' logic, resulting in an incoherent, nonsensical jumble of action.

In any of these cases, the resulting agent will display the symptomatology of schizophrenia as I defined it earlier. My conclusion is that *schizophrenia is a direct and inevitable result of atomization*. It is a fundamental property of our agent design strategy.

The catch-22

If schizophrenia is caused by atomization, then it would seem that the most obvious way to get rid of it would be to get rid of atomization. This is, in fact, the agent design strategy proposed by Loyall: all the parts of the agent should be designed together. But this solution is impractical for large, complex agents, as we discovered when we built the Woggles. Atomization is an essential strategy for simplifying phenomena enough that we can understand them. Getting rid of atomization means understandable, modularized code is thrown out the window. Making behaviors arbitrarily complex and interrelated also makes them arbitrarily difficult to debug and comprehend. For the sake of being able to program, we need a certain amount of atomization.

At this point we are backed into a corner. The final conclusion of the arguments made here is that atomization causes schizophrenia, but we need atomization to write code. This is a vicious circle.

If this argument holds, then schizophrenia will not be solved by a clever new algorithm. It then represents the absolute limit point of current ways of understanding agents. As far as I can tell, *schizophrenia cannot be addressed within current AI frameworks*. It is a dead end.

The goal of this thesis is to change this. I believe AI can and should be done differently. This will require us to rethink the foundations of AI. Such rethinking has traditionally been done through importations from the sciences and analytic philosophy. While many of these importations have been ingenious, insightful, and stimulating, I suspect they are not enough, since they generally share the same atomistic principles. To solve schizophrenia, I believe we will need a radically new perspective. This thesis explores the possibility of getting that perspective by understanding AI as culture.

In the next chapter, we will use this humanistic approach to come to a deeper understanding of schizophrenia and its relationship with intentionality. Cultural theory and antipsychiatry make some suggestions about how schizophrenia can be handled. This different way of handling schizophrenia will become the basis for the technical results in the second half of the thesis.

Chapter 3

Schizophrenia, Industrialization, and Institutionalization

This chapter, and indeed this entire thesis, has its origin in a week in 1991, which I had the pleasure of spending hospitalized for depression at Western Psychiatric Institute and Clinic in Pittsburgh. This week was a turning point in my life, not because of any particular help I received in my hour of need, but because, for the first and hopefully last time in my life, I discovered what it was like to spend 24 hours a day under the surveillance of a scientific system that tries to regulate every aspect of individual, human, subjective experience. The amazing paradox that became clear over my days in this fishbowl environment is that the more subjective experience is placed under surveillance, categorization, and attempted control, the less it is actually observed, understood, and influenced. Using a label like 'atypical personality disorder' and writing a prescription for Anafranil did little to address the existential crisis that had brought someone to this unbearably painful point in life. Instead, it tended to build barriers, to separate patient from doctor, to make the doctor feel competent to judge the humans in his or her care as instances of a category, and to keep the doctor from being drawn into the muddled details of treating him or her as a complex, messy, fellow being.

This experience made clear for me the limitations of objective approaches to understanding subjective experience. Certainly, objectivist[1] knowledge traditions such as psychopharmacology have their place; no one can deny, for example, the power of lithium to give manic-depressives stability in their lives. At the same time, these objectivist traditions, particularly when seen as the only and ideal goal of all human intellectual endeavor, leave something important out: individual human experience, with all its rich and ambiguous implications, with its meanings not objectively available but to be sorted out moment-by-moment by specific

[1] By 'objectivist' I mean "having the goal of objectivity." Whether these forms of knowledge production can ever actually achieve true objectivity (whatever that is) is far beyond the scope of this thesis, though my guess would be that, since they are so bound up in interpersonal relations and cultural traditions of what is and is not normal, they probably can't.

people in concrete situations, experience which is most adequately jointly approached not as doctor-patient, not as subject-object, but as equal and unique but related individuals leading complex and very real lives.

In this chapter, we will look at the limitations of objective knowledge and possibilities for subjective understandings of human life from various perspectives. The goal is to flesh out the understanding of schizophrenia and its relationship to atomization that we started exploring in Chapter 2. We will do this by looking at two particular case studies that relate to AI's ways of understanding agents: industrialized understanding of workers on the assembly line and psychiatric understanding of mental patients. In both of these cases, we will see similar processes at work: the attempted categorization and control of sometimes frighteningly ineffable human behavior through the application of objectivist forms of knowledge production. In each of these cases we will also find limitations in the extent to which these techniques can actually be used to control and understand human behavior; subjectivist critiques of each tradition will provide alternative ways of understanding that may be more helpful.

Dear technical reader, this chapter may be quite a challenge to you (though it may also be an unaccustomed treat). Although the conclusions of this chapter will form the basis for the technical results in the rest of the thesis, little that I say in the pages of this chapter will bear directly to problems in autonomous agents. The approach in this chapter will be exclusively humanistic. The form of argumentation will be largely metaphorical; I will try to draw out metaphorical connections between various cultural practices that relate to schizophrenia in autonomous agents. There is a logical argument to be found here, but in the rhetorical forms of cultural theory, the 'point' is not only the argument but also the details of the concrete situations that are looked at. "What matters above all is not to reduce everything to a logical skeleton, but to enrich it, to let one link lead to the next, to follow real trails, social implications" ([Guattari, 1984], 259). Fundamentally, this chapter is selling, not an argument, but a particular and rich way of seeing which can apply to various parts of life, including but not limited to AI. Good luck and enjoy!

Case Study 1: AI as Industrialization

When I took my first course in AI, I was gripped by an intense fascination: why on earth would anyone think AI was even possible? With all the amazing, strange, wonderful, horrible, bizarre things that humans do and are, what would ever possess anyone to think that this miraculous existence could be reproduced by a machine? How can indefinable, ungraspable consciousness be thought to be 'implemented' in machinery, apparently as a set of search algorithms? What kind of strange and twisted view of humanity is embodied in chess-playing machines as philosophy of life?

After being immersed in the field for several years, I began to see AI as a natural view on life; it becomes hard to remember this initial sense of wonder. In fact, AI researchers tend to feel that any mention of it is a sign that the bearer of the wonder is either a fuzzy-headed believer in the supernatural or suffering from a little heat stroke that a good nap might cure. But even if we are willing to grant AI an intellectual certainty, a question still lurks: why is it that, at this moment in history, AI as an intellectual endeavor seems to so many not only possible but self-evident? What is it about our current way of thinking that makes the very idea of AI a natural extension of our intellectual traditions?

In and of itself, the idea of reproducing life is nothing new; medieval Jews already had the tradition of the Golem, an effigy magically brought to life; the 19th century brought the organic horror of Frankenstein. But the techniques by which AI approaches the problem of creating artificial life are different; the creation of life is no longer a question of magic, alchemy, or biology, but one of information. Artificial beings are not made of clay or rotting body parts, but of algorithms glued over bits of silicon and robotic machinery. AI and cognitive science approach life not as a mysterious spiritual or biological process to be engaged in or mimicked but as a machine, like any other, to be designed and controlled. AI, in this sense, is the next step of industrialization: having replaced worker's bodies with robotic machinery, we are now developing replacements for the worker's minds. In this section, we will look at AI as the industrial mechanization of subjectivity. We will dig deeply into industrialization to understand AI's inheritance from it: techniques, philosophies, but also problems, among them schizophrenia.

> "In a sense, the mechanical intelligence provided by computers is the quintessential phenomenon of capitalism. To replace human judgement with mechanical judgement — to record and codify the logic by which rational, profit-maximizing decisions are made — manifests the process that distinguishes capitalism: the rationalization and mechanization of productive processes in the pursuit of profit.... The modern world has reached the point where industrialisation is being directed squarely at the human intellect."([Kennedy, 1989], 6

Industrialization as Mechanization

The history of the Industrial Revolution is, among other things, a story of the gradual replacement of workers by machines. This fable proceeds as follows: in the beginning there were craftspeople, who owned their own tools, who manufactured articles in their own, idiosyncratic ways, whose work was largely integrated with their way of living. As the Industrial Revolution begins, these workers begin to be collected into factories, where they work together using the owner's tools. This owner, in an attempt to make work more efficient, begins to streamline the production process. Instead of having each worker build a piece from beginning to end, the production line is developed, where each worker works on some small part of the final piece. Work is broken up into stages, each of which is accomplished by a single worker; each stage is standardized so that articles can move from stage to stage without breaking work rhythm. Once work is divided up into standard stages, some of the steps can be done by a machine. Instead of building an article from beginning to end, workers now tend machines which are each doing small steps of the article's production.

At each stage, work becomes more rationalized, predictable, and efficient. Workers on the assembly line can generate more articles, and the articles lack the idiosyncratic variation of normal craftwork. Instead of doing whatever he or she wants in a haphazard order, a worker has a fixed set of steps he or she engages in. The intelligence of the worker, which s/he previously needed in order to monitor what s/he was doing and make active decisions about how work should proceed, is now embodied in the structure of the assembly line. Workers no longer need to think; the factory machinery does the thinking for them. Even before computers, industrialization takes the first baby steps of AI.

These traces of industrialization can be seen in the way we build agents today. The AI researcher building an agent follows the same basic line as the factory manager designing new production processes. Just as the factory manager attempts to design and reproduce a pre-existing work process, the AI researcher would like to copy a natural process — an agent or idea of an agent. Just as the factory manager breaks this process into logical steps, figuring out which steps should happen and in which order, AI researchers analyze the agent's behavior, to categorize its activity into typical behaviors and to enumerate the conditions under which those behaviors are appropriate. And just as the factory manager embodies each step in machinery which can run with a minimum of human supervision, AI designers implement a mechanical version of each behavior, hooking them together so that they largely reproduce the imagined or real behavioral dynamics of the original creature. The early industrialist and the AI researcher are engaged in the same project: we analyze, rationalize, and reproduce natural behavior.

Marxism: a Demystification
Since the beginnings of the Cold War, Marxism has had a bad name in the United States. Its use by Communist totalitarian systems has not done much for its image in the public eye. Since the fall of the Soviet Union, much of the educated public has been led to believe that Marxism, like Communism outside of China and Cuba, is dead.
But Marxism is not just a (seemingly failed) political doctrine predicting the end of capitalism, but also a thriving intellectual tradition. This tradition includes some of the greatest thinkers of the 19th and 20th centuries, most notably but certainly not limited to Marx himself, an intellectual and scholar whose influence has been felt around the globe and in many disciplines for a century and a half. In this tradition, Marxism can be understood simply as a theory of specifically industrial society; this face of Marx forms the basis not only of scary left-wing political theory, but also of much of modern economics.
It is impossible to do any serious analysis of industrial culture without Marxism, and this thesis is no exception. This means a continuing dialogue with the Marxist tradition, one that takes into account not only Marx's original writings, but also contemporary reinterpretations of them. Most notably, it seems that Marx's prediction that capitalism would lead to its own downfall is probably incorrect; and nearly all Marxist-influenced thinkers consider the reduction of all cultural activity to class warfare as long since passé. Nevertheless, that still leaves plenty of grist for the intellectual mill. Here, I will focus particularly on Marxist analysis of the changing experience of being human as more and more kinds of labor become mechanized.

At first blush, a difference between AI research and industrialization may seem to be that AI seeks to reproduce intelligence, whereas the industrialist is not so much interested in reproducing work processes as in maximizing profit on their output. This means that post-industrial work is radically different from pre-industrial work, in both the qualities of the articles produced, and in the human experience of engaging in that work. The very act of embodying work in the production line changes the nature of work. Work becomes more rationalized and less personal; workers are more dependable and more bored; the articles produced become more standardized and less individual.

But in Chapter 2, we saw that, just like early industrialists, AI researchers do not create absolutely faithful reproductions of the living beings they seek to emulate. We looked at the special properties that artificial creatures have when they are built using atomized processes. Atomization, we learned, introduces its own qualities that can be recognized in the creatures generated with it, among them schizophrenia.

Similarly, several special qualities of post-industrial work and life have been identified by cultural theorists and industrial historians:

- *Reification* — Things that were once thought of as ineffable or abstract become thought of as concrete. 'Labor,' for example, which was once not strictly separated from the rest of life, becomes something that is measured and sold per piece or per hour. Once things are reified, they can be sold, becoming *commodified*, to be exchanged for particular sums of money.

- *Specialization* — Workers no longer engage in the entire work process; rather, they each perform some small function within the process as a whole. Without an overview of the process, workers no longer need or are able to adapt to one another; each part takes place without reference to the others. Without feedback between the pieces, each piece is built in isolation, the whole then being merely the sum of each individual, separately designed atomic part.

- *Atomization* — The production process, which was once a wholistic attribute of individual workers, is broken into rationalized parts, each of which is embodied in pieces of machinery or in production rules that regulate how they interact. Workers, who were once thought of as individual humans deeply embedded in the context of their daily life, now become interchangeable parts of the production process, whose time is to be sold to the highest bidder. They move from factory to factory, no longer connected to their home place or even to a particular manufacturer. Workers see themselves as free and atomic individuals, bound by no human ties.

- *Standardization* — The idiosyncrasies of craftwork means that one can never be sure what the produced goods will be like. The factory owner, on the other hand, who consolidates craftwork and has promised broader distribution networks goods of a particular kind and quality, wants to have some guarantees that the factory will produce similar goods no matter which workers are present on a particular day. The idiosyncrasies of personal work are no longer valuable; instead, the owner introduces steps of production control to make sure that the output is always similar. The qualitative, human, individual dimension of work is eliminated, replaced by efficient, controlled, and standardized work processes.

- *Formalization* — The individual, material properties of workers and the material they operate on is ignored, except insofar as it impinges on the production process. As the production process becomes more and more efficient, extrinsic considerations — whether social, spiritual, or physical — are left out. The production manager thinks of the production process in terms of abstract steps, without reference to the particular identity of the worker or chunk of material involved; the factory is set up to enforce this abstract, impersonal view, which then seems to be an accurate reflection of the real. Individual differences become 'noise,' unvalued and only reflected upon in order to control their effects.

- *Mechanization* — In order to maintain standard production, workers are given less and less leeway in decisions about their jobs. Rather than relying on the worker's judgment, the factory manager uses standardized production rules to ascertain that the product is always made the same way. As the steps of the production process are more and more formalized, the worker's intelligence becomes less and less pertinent. Once the worker's intelligence is no longer needed, the worker can be replaced by a machine.

These trends in industrial culture are rooted in factory work, but they did not stay confined to the factory for long. Workers, who spend a large portion of their waking hours interacting with machinery on the production line, take home the values that that system has ground into their bodies. Production line designers, factory owners, and managers, spending their days designing machinery and optimal control of the human-machine interface, do not always forget their machinic view on life on their days off.

More insidiously, the drive of the assembly line, powered by the money its efficiencies can bring, spreads into other intellectual fields: factory owners hire inventors, scientists, and engineers to design machinery and the production processes they support; they hire social scientists and management experts to design worker compliance (Total Quality Management is born). Each of these fields, applying itself within the context of the assembly line, starts to find more and more ways to generate interesting results within the factory work framework, slowly and mostly unconsciously taking over the assumptions of formalization, atomization, and so on that that framework presupposes. Factory owners lobby for laws that support and reflect their point of view on factory work. Both private and public institutions are set up to explicitly market these practices; the military, for example, played a large hand in encouraging development of standardized manufacturing [DeLanda, 1991]. Other businesses, which are not strictly factory-oriented, envy the efficiency and rationality of factory work, and begin to apply some of their ideas to improve their own processes. Soon the countryside is dotted with identical, standardized, efficient fast food restaurants with its teenaged automatons taking on the role of factory machinery; no one living in these cultures can escape the force of industrialization, even on their lunch break.

All this talk of workers in the factory and the assembly line may come across as antiquated today. How many of us are still factory workers on the assembly line? When ‘us’ means the readers of this thesis, the answer must be very few. After all, we late 20th century Westerners are no longer in the industrial era; we are brave new citizens of the Information Age![2]

But the forces of industrialization, far from having disappeared, have become so ingrained in our daily lives that they are taken for granted. If you live in the West, and especially if you are American, industrialization is the air you breathe and the prepackaged food you eat. Your life becomes more and more mechanized as your bank teller is replaced by an Automated Teller Machine, your receptionist is replaced by a voice mail program, the telephone solicitor who has interrupted your dinner every night for the last 6 years is replaced by an auto-dialer with a cheery robotic recording. The last bastions of your craftwork slowly give way as universities become digital diploma mills, offering impersonal and standardized distance learning to students who are no longer limited by the bonds of location or social interaction [Noble, 1998]. When in Paris, you eat your standardized lunch at McDonald’s, knowing that, while it may not be very good, it won’t expose you to the idiosyncrasies of local cuisine — any calf brains will be ground beyond recognition into your Big Mac. You reify yourself as you sell 4 hours a day to each of 3 part-time jobs, trying to still maintain a full sanctified hour of Quality Time with your youngster — go ahead, sell yourself until you can afford to buy yourself back! You have specialized yourself, become the world expert on polynomial kernel support vector machine with fractional degree, unable to discuss your work with more than 3 or 4 colleagues because it is so hopelessly obscure (but nevertheless breathtakingly important). You atomize yourself, cut off from your extended family, perhaps even from your spouse and children, moving every 7 years in an evanescent search for the better life. How many times a day do you formalize yourself, jacking into cyberspace to become blissfully unaware of the constraints of your undeniably material, geographically located, and mortal flesh — at least until your RSI kicks in?

[2]Note that life looks very different to those in the third world, for whom underpaid and dangerous work on the assembly line is still a very real daily experience.

No, industrial culture is not confined to the 19th-century factory; it continues to live itself through us on a day-to-day basis. Industrial culture is not just an attribute of a now marginal work-life; it colors nearly every aspect of late 20th century Western existence [Strasser, 1982]. It is not just a way of producing goods, but a new and not always positive way of being. We are post-industrial humanity: reified, specialized, atomized, standardized, formalized, mechanized; we are nonstandard flesh, the weak link in a network of machines.

"[T]oday Western man has become mechanized, routinized, made comfortable as an object..." ([Josephson and Josephson, 1962], 10)

Taylorism and Schizophrenia

What is it like to live a post-industrial existence? For humans, industrialization is often an experience of being more and more dominated by systems of machinery, of both the technical and bureaucratic kinds. This is certainly the case for craftworkers, whose work historically consisted of skilled tinkering in the workshops of their houses, but presently generally involves the monitoring of raw materials as they are fed through massive machinery. Rather than applying their intelligence and skill to an ever-renewed activity, taking pride in the result of their handiwork, workers go through repetitive and mindless motions that are stipulated from beginning to end by the production handbook in order to create finished products they will never see. The experience of being a worker was once work; now, it is being an appendage to a machine.

"Taylorist man is a slave to the movements of a machine, and he cannot control it either technically or socially. Above all, he suffers from the divorce between that part of his body which has been instrumentalized and calibrated and the remainder of his living personality." ([Doray, 1988], 82)

In a sense, it is workers themselves who have become mechanized. Georg Lukács has shown that the mechanization of the work process does not stop with production itself; the workers themselves are progressively designed and controlled as machines.

> If we follow the path taken by labour in its development from the handicraft via co-operation and manufacture to machine industry we can see a continuous trend towards greater rationalisation, the progressive elimination of the qualitative, human and individual attributes of the worker. On the one hand, the process of labour is progressively broken down into abstract, rational, specialized operations so that the worker loses contact with the finished product and his work is reduced to the mechanical repetition of a specialised set of actions. On the other hand, the period of time necessary for work to be accomplished (which forms the basis of rational calculation) is converted, as mechanisation and rationalisation are intensified, from a merely empirical average figure to an objectively calculable work-stint that confronts the worker as a fixed and established reality. 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 specialised rational systems and their reduction to statistically viable concepts. ([Lukács, 1971], 88)

Taylorism, or scientific management, is the apogee of this view of worker-as-machine. The goal of Frederick Taylor's scientific management is to increase the efficiency of work processes by analyzing and optimizing not only the machinery itself, but also the way in which the worker uses the machines. Through time and motion studies, the worker's motions are examined; all motions are forbidden. Rather than letting workers interact with machinery in any way that they saw fit, Taylorists determine the "one best way," the most efficient possible use of the machinery. Upon Taylorization, workers are generally given detailed instructions of every movement they should use to accomplish their job. Nothing is left to chance; nothing is left to worker ingenuity; nothing ever varies. With Taylorism, the rationalization of the work process, having extended into the worker's mind, is complete.

"It is not simply status-hunger that makes a man hate work that is mindless, endless, stupefying, sweaty, filthy, noisy, exhausting, insecure in its prospects and practically without hope of advancement. The plain truth is that factory work is degrading" ([Swados, 1962],111).

Despite the radical successes of Taylorism in improving the efficiency of industrial work, it also has some unexpected negative effects. Taylor thought that workers would be happy to be able to work more efficiently and make more money. Instead, unions object to Taylorist techniques because they reduced workers to mindless objects, ignoring the expertise of skilled workers in favor of scientific analyses by outside experts. Workers find the absolute banalization of the work process that Taylorism implies unbearable; Taylorist work is both repetitive and mindless, on the one hand wearing out workers' bodies with repetitive stress injuries, on the other boring them senseless [Doray, 1988].

Taylorism demands that, not only the process of production, but humans themselves become rationalized and mechanized. The difficulty in this plan is that people are not machines. While Taylorists are able to categorize and optimize worker movements, they do so while ignoring the worker as human being. The result is that a small part of the worker's existence is identified and reinforced; the remainder is repressed, until ignored aspects demand attention when the worker is injured, becomes distracted, or simply refuses to submit to such a repressive regime (or, in the case of postal workers, shoot their co-workers and bosses).

Ironically, while Taylorism leaves something to be desired for its original goals, it is extremely well-suited as a basis for Artificial Intelligence. While workers cannot handle these repetitive, mindless activities, they are perfect for robots. Numerous scholars have pointed out that Taylorism is the last intellectual stop before AI: as soon as work is reduced to mindless, rote movement, idiosyncratic and moody human workers can be replaced by controllable and indefatigable robots, removing the last unpredictable part of the production process.

The principles of Taylor — quantifying and rationalizing human behavior, reducing intelligent behavior to a set of independent, predictable, and interchangeable parts, removing all traces of human idiosyncrasy, creativity, and craftwork — are now suspect in management, but live on in the engineering tradition in computer science. Michael Mahoney, a historian of science, points out with some surprise that in software engineering, the arguments are not about whether the principles of Taylor are correct, but about how to apply them [Mahoney, 1997]. This observation extends to Artificial Intelligence — in many ways, AI is simply the late 20th-century reincarnation of turn-of-the-century traditions of human engineering and control.

In Taylorism, as in agent design, a coherent and wholistic behavioral dynamic is carved into chunks. Individual pieces of behavior are identified and rationalized. In Chapter 2, we noted that this partial rationalization leads to schizophrenia, and the same effect happens under Taylorism. Again, this schizophrenia is not meant as a psychiatric label (although it certainly seems possible that assembly line work could drive someone insane); by schizophrenia I mean a disintegration of subjective experience as some parts of a person are brought out and others repressed. This schizophrenia is experienced directly as the boredom, degradation, and depersonalization of assembly line work.

Schizophrenia for the Masses: Industrialized Life

"Our society produces schizos the same way it produces Prell shampoo or Ford cars, the only difference being that the schizos are not salable." ([Deleuze and Guattari, 1977], 245)

This form of schizophrenia is not simply a result of Taylorization, although certainly Taylorism displays it more extremely. Marxist and post-Marxist scholars understand this kind of schizophrenia to be a result of simply living in industrial society [Lukács, 1971] [Deleuze and Guattari, 1977]. This is because all of us in post-industrial society — whether assembly line workers, hamburger flippers, or university professors — are constantly coming into contact with machinery and bureaucracy that is set up to ignore most of what we might value in ourselves. We are enmeshed in objective 'laws,' imposed from outside: the rules of the assembly line, the invisible hand of the market, the laws of physics. We live our lives qualitatively, while continuously being asked to make decisions and define ourselves in terms of quantitative and inhuman systems. These mindless systems come set up with a priori categories; our freedom and humanity is manifested only in that we can choose which category we want to be processed through. The industrialized doctrine of individuality is "choose 1 of *n*:" you can choose one of 6 Extra Value Meals, drive your sport utility vehicle to one of 14 suburban malls, click on one of 8 links, buy one of 123 kinds of cereal, punch in one of 9 responses to the voice-mail prompt, vote for one of 3 politicians, identify with one of 4 ethnic groups; but if you want to stay in the system you cannot meaningfully choose 'none of the above,' or, God forbid, half of one and a third of another with a little bit of something extraneous mixed in.

George Ritzer studies the extent to which themes from assembly line work, Taylorism, and bureaucracy have infiltrated our daily lives [Ritzer, 1993]. This 'McDonaldization' of society is based on the growing cultural importance of *formal rationality*, i.e. systems under which people try to find the best ways to achieve pre-given and unquestioned goals, not according to their personal feeling for how it should be done, but by reference to formal systems of rules and regulations. This kind of rationality is interested, like Taylorism, in "the one best way" to do things, and is suspicious of the ability of individual people to judge things for themselves. Instead of leaving decisions up to the people involved, technical, legal, and bureaucratic systems are set up so that the 'best' way to do things is also the natural way. In industrialized society, 'best' is judged along four axes:

- *efficiency* — The production line maximizes the efficiency of craftwork; everything extraneous to optimal performance is removed, including personal idiosyncrasy and the joy of handiwork. For industrial culture, the number one goal is to satisfy needs quickly and without waste. Rather than lingering over a satisfying meal, the goal is to get customers in and out as quickly as possible. Why waste valuable time cooking a meal from scratch, when a frozen prepackaged pot pie is cheap and oh, so easy?

- *quantifiability*— In order to maximize efficiency, engineers calculate as much of the work process as possible: piece rates, materielle usage, worker movements, labor costs, recidivism rates. Things that cannot be calculated are devalued. Cost/benefit analyses weigh quality of life against cold, hard, calculable cash. The soul can't be weighed, so it must not exist.

  Quantifiability implies that more is better. The chain with the most stores must be the best. We buy, not the best-tasting sandwich, but the one with the largest pile of unidentifiable ground meat. Airlines advertise, not "We have the most pleasant flights," but "We fly to the most cities." The more you buy, the more you save!

- *predictability* — One of the major advantages of assembly line work is that the output of the assembly line is predictable. The phalanx of cars come marching off the assembly line, each exactly identical, with interchangeable parts, each with the same new car smell, the same ride, the same fluffy upholstery, the same engineering mistakes. Predictability substitutes for familiarity: wherever we go, the Days Inn is exactly the same, with the same cheerful desk clerks telling you to have a nice day, and the same style of insipid sit-com grinding out its laugh track from the satellite TV in your room. On your bus tour of Europe, there are no unpleasant surprises: 1 day per city, carefully sanitized local color, and the natives you meet all speak perfect English.

- *control* — Life (and in particular human behavior) is in many ways not inherently predictable, so the holy grail of predictability is only achievable through the hefty use of controls. Unpredictable humans are replaced and controlled by technology and bureaucracy: the ATM never miscounts, the computerized tram keeps its doors open for exactly 5.3 seconds, and the fast food worker does not get a chance to misspeak while regurgitating the manual's "Fries with that?" The customer can remain cost-effectively always right when s/he only has a choice of 5 menu items, and the high-intensity fluorescent lighting chases him or her out of the restaurant before s/he becomes an economic liability.

Each of these values certainly has its place. Inefficiency, incalculability, unpredictability, and lack of control are clearly not particularly preferable to their opposites. But Ritzer points out that under formal rationality, each of these values is elevated to an absolute. And when rationality is pushed too far, the result is, paradoxically, irrationality. A cheap fast food meal with huge portions, wolfed down in 5 minutes, is not necessarily better than a less 'efficient' home-cooked meal with quality ingredients. A packaged group tour with all activities carefully homogenized and isolated is safer, but not necessarily better, than a vacation requiring true contact with alien cultures and experiences. America has certainly pushed the envelope of homogenized, commodity-laden, safe, and predictable existence, but whether we truly maintain quality of life is an open question in a nation of the obese, who fuel the purchase of all the latest high tech fantasies by road raging across miles of asphalt from the suburbs to a 10 hour work day, then crawl back home to frozen dinners consumed silently in front of the stereo, big-screen high-definition TV. Even Taylorism, which subsumes all human values to the goal of efficiency, is inefficient in the sense that, by reducing work to repetitive motion, it wastes the worker's talents and judgment.

Under formal rationality, that which can be predicted and controlled is analyzed and rationalized. That which does not fit into those systems is ignored or denigrated. The result is the atomization, the fragmentation, the schizophrenization of daily life: the mindless suburban utopia as seen on TV masking violent death in the inner city; the back-to-nature marketing of enormous, environmentally destructive vehicles that will spend their lifespans only on urban highways and shuttling teenagers to and from the strip mall; taboos on sex mixing with advertising based largely on sex; unthinking bible-thumping TV evangelism providing hedges for the utter vacuity of spiritual values in public discourse. Little rationalities we surely have — the world's greatest can opener — but no sense as to how they should fit together into a meaningful life. Meaning itself — being one of those old-fashioned, inadequately calculable terms — no longer matters. Life, like our wetlands, is drained dry and replaced with a Wal-Mart.

"In milling and baking, bread is deprived of any taste whatever and of all vitamins. Some of the vitamins are then added again (taste is provided by advertising). Quite similarly with all mass-produced articles. They can no more express the individual taste of producers than that of consumers. They become impersonal objects, however pseudo-personalized. Producers and consumers go through the mass production mill to come out homogenized and de-characterized — only it does not seem possible to reinject the individualities which have been ground out, the way the vitamins are added to enrich bread." ([Van den Haag, 1962], 183)

The law of industrialized culture is 'choose one of *n*.' Categories which were once abstract and qualitative are reified, making them quantitative and strictly delimited. Human qualities — your time, your work, your body — become commodities to be sold at will. Intelligence becomes IQ, family traits become genetic predispositions, class becomes income level, existential anxiety becomes a mental disease with its own number in the *Diagnostic and Statistical Manual of Mental Disorders*. In the process two things happen: definitions of these categories become so strict that they exclude much of what we find valuable in their informal counterparts; and in the process of setting up strict and delimited categories we lose the interrelationships between them. So, for example, in AI behavior is no longer a wholistic style of activity through which a being's existence is expressed; it becomes a set of atomically defined and separately written behaviors, of which industrialized agents choose one of *n*. Industrialization involves the loss of wholism and interconnections in favor of individually rationalized and atomically related parts; it leads to schizophrenia, the fragmentation of subjective experience itself.

Industrialized Science

The rise of industrialization has been accompanied by a rise in the importance of science and engineering. This link is not accidental. Science and technology give industrialists the ability to predict and control processes, providing the motive power for industrialization. At the same time, the industrialized world view is sympathetic to the scientific assumption that life is fundamentally a mechanical process that can be understood and controlled. Industrialization provides funding for the parts of science that are particularly useful for it, reinforcing those styles of science at the expense of others. Science in turn provides industrialization with

"Machines — and machines alone — completely met the requirements of the new scientific method and point of view: they fulfilled the definition of 'reality' far more perfectly than living organisms. And once the mechanical world-picture was established, machines could thrive and multiply and dominate existence." ([Mumford, 1934], 51) rationales for its activities. Science and industry become symbiotic, each reflecting aspects of the other.

Artificial Intelligence is no exception. From planning and scheduling of shop activities to robots for the assembly line, to reinforcement learning for process control, to automatic translation of manuals for equipment assembly, AI works on the problems of industrialization and, in turn, imbibes its values. Efficiency; quantifiability; predictability; control: Ritzer's values of industrialization are also the values of AI. They can be seen in the view of intelligence in AI, so different from most people's day-to-day experience of existence: the calls for rational, goal-seeking, provably correct agents, working efficiently to solve problems. They are reflected in the fundamental hope of AI: that most if not all of human behavior can be rationally analyzed, quantified, and reduced to algorithms reproducible in the machine.

AI is not alone; it represents in miniature the themes of post-industrial science, themes which are inherited from industrialization.

- *Reification* — Science works by approaching the multitude of phenomena of existence to find ways of sorting and categorizing them into well-defined categories. Animals are categorized into species, pain and discomfort into diseases, activity into behaviors. While the categories are always subject to revision, this involves the replacement of one kind of rigid definition with another, not the wholesale dissolution of categories. Classification is essential to science; without it, regularities cannot be discovered [Kirk and Kutchins, 1992].

- *Specialization* — Modern science has become more and more specialized and esoteric. Science is split into many heterogeneous sciences, each studying its own phenomena in its own way. It is not even clear how to relate the subfields of a particular discipline, let alone how biology, chemistry, physics, psychology, and sociology can be combined to form one consistent world view. In this sense, science, like modern consciousness, is fragmented and incoherent.

- *Atomization* — The methods of science involve breaking up phenomena into subparts, studying these parts in isolation, and trying to reconstruct the full phenomena from these presumably independent parts. "[T]he ideology of modern science... makes the atom or individual the causal source of all the properties of larger collections. It prescribes a way of studying the world, which is to cut it up into the individual bits that cause it and to study the properties of these isolated bits" ([Lewontin, 1991], 12-13). Lewontin points out that this way of conceptualizing the world, which comes to us in post-industrial society so naturally, would have been unthinkable in the Middle Ages, when nature was seen as essentially wholistic; dissecting nature was thought to destroy its essence. But when all of society is thought to consist of atomic, free, and independent individuals, it is not so strange to think of nature this way, too.

- *Standardization* — Science understands all electrons, all anxiety disorders, all elephants as basically alike. Yes, there are individual differences within a category, but the very construction of a category implies for scientists that the rules regarding that category apply to its members in the same way. Under industrialization, "the human qualities and idiosyncrasies of the worker appear increasingly as *mere sources of error* when contrasted with these abstract special laws functioning according to rational predictions"([Lukács, 1971], 89). Similarly, any variation of the behavior of scientifically classified objects from the norm is considered statistically manageable 'noise.'

- *Formalization* — Science differs from alchemy in that the individual, material, idiosyncratic attributes of objects are considered unimportant. Rather, the ultimate goal of science is the reduction of the material to mathematics. Truly elegant scientific theories represent complex reality in terms of a few simple, well-defined laws, formal representations into which scientific objects can be plugged with the minimal possible reference to their idiosyncratic individuality.

  For the same reason, the context of scientific work is often minimized or forgotten. The scientist reduces not only the idiosyncrasies of the scientific object, but tries to remove his or her own idiosyncrasies as individual from the results of scientific work. "[S]cientific experiment is contemplation at its purest. The experimenter creates an artificial, abstract milieu in order to be able to *observe* undisturbed the untrammelled workings of the laws under examination, eliminating all irrational factors both of the subject and the object. He strives as far as possible to reduce the material substratum of his observation to the purely rational 'product', to the 'intelligible matter' of mathematics" ([Lukács, 1971], 132).

- *Mechanization* — The scientific worldview is a mechanical worldview. References to the 'soul,' to God, to the unknowable, to the very possibility of free will, which might be considered signs of a healthy respect for the limits of human ways of knowing, instead are considered highly suspect and even laughable. Instead, one of the ultimate goals of scientific knowledge is the synthesis of physics, biology, and psychology, into a complete description of human beings as fully mechanical systems. The body is a machine; with the development of cognitive science, the mind is a machine as well. This mechanistic viewpoint is not seen as metaphor, but as reality: Lewontin notes that while in Descartes' day, the world was considered to be *like* a machine, in our post-industrial existence we really consider the world to *be* a machine [Lewontin, 1991].

Post-industrial science works on the theory of the assembly line: "the process as a whole is examined objectively, in itself, that is to say, without regard to the question of its execution by human hands, it is analysed into its constituent phases; and the problem, how to execute each detail process, and bind them all into a whole, is solved by the aid of machines, chemistry, &c." ([Marx, 1967], 380). Like industrial engineering, science understands life by decontextualizing and dissecting it — taking it apart, analyzing each part separately, and then combining these independent forms of understanding into a functional but nevertheless fragmented whole. "Scientific... 'good sense' operates in essentially the same way as common sense: isolation of the typical individual (considered outside the real flow of its actions; as essentially dead); decomposition into parts and determination of intrinsic qualities (dissection); logical recomposition into an organic whole exhibiting signs of 'life' (artificial resuscitation)..." ([Massumi, 1992], 97). The wholism of science is a summation of individual, independent parts, each rationalized separately, each acting without reference to the others: Lukács's "objective synthesis" (88), the sum of calculation, arbitrarily connected.

The result of this process of fragmentation is schizophrenia. The object of study in science is split into a thousand pieces, each of which is rationalized separately and reunited in a parody of wholism. The union of these parts is incoherent; they may fit together in places, but only by accident; their necessary connections were left behind at the moment of dissection. And all that is not amenable to rational analysis is also left behind, forming a residue of noise that marks the limit-point of rationality. Schizophrenia is the uncategorizable; in the feedback loop between rationality and incoherence, schizophrenia is the short-circuit.

Case Study 2: AI as Institutionalization

"A man who says that men are machines may be a great scientist. A man who says he *is* a machine is 'depersonalized' in psychiatric jargon." ([Laing, 1960], 12)

So far, schizophrenia has functioned as an abstract term in this thesis, a breakdown in overall cohesion that comes about when life is micro-rationalized. However, schizophrenia is not simply a trendy theoretical term, but also a lived reality; and there are important relationships between AI as an intellectual discipline and the experience of being a schizophrenic person, especially as understood by institutional psychiatry.

In particular, the flip side of the AI concept of consciousness-as-machine is the schizophrenic experience of self-as-machine. Critics of institutional psychiatry argue that this 'delusion' (or, better put, unique and painful existential position) is reinforced under a scientific psychiatry that attempts to explain schizophrenia in mechanistic terms. Taking an objective perspective on schizophrenia, seeing patients' behavior not as an expression of their unique selves but as mere symptomatology of a disease, fundamentally involves denying those patients' already marginal experience of personhood, rendering schizophrenics incomprehensible, their speech no more than word salad. Institutional psychiatry, by objectivizing the schizophrenic and schizophrenia, splits the schizophrenic from his or her context and from his or her disease, repeating the fragmentation of subjective experience that is a hallmark of schizophrenia. These moves parallel the decontextualization, reification, and fragmentation of behavior that occurs in AI. In this section, we will look at these problems in institutional psychiatry in more detail; proposed solutions to the problems inherent in this mechanization of patient psychology will become the basis for rethinking AI in Chapter 4.

Institutionalization as Mechanization

"Toute l'histoire de la folie... n'est que la description de l'automatisme psychologique." ([Janet, 1889], 478)

In the late 1800's, Pierre Janet identified one of the more baffling symptoms of schizophrenia — the *sentimente d'automatisme*, or subjective experience of being a machine. This feeling is the flip side of AI's hoped-for machinic experience of being subjective, and is described by one patient this way: " 'I am unable to give an account of what I really do, everything is mechanical in me and is done unconsciously. I am nothing but a machine' " (an anonymous schizophrenic patient; cited in ([Ronell, 1989], 118)). R. D. Laing describes how some schizophrenic patients experience or fear experiencing themselves as things, as its, instead of as people [Laing, 1960]. Schizophrenia is, for some, a frightening feeling of being drained of life, of being reduced to a robot or automaton.

This feeling of mechanicity is correlated with a fragmentation of the affected patient's being; sometimes, a schizophrenic patient's very subjectivity seems to be split apart.

> In listening to Julie, it was often as though one were doing group psychotherapy with the one patient. Thus I was confronted with a babble or jumble of quite disparate attitudes, feelings, expressions of impulse. The patient's intonations, gestures, mannerisms, changed their character from moment to moment. One may begin to recognize patches of speech, or fragments of behaviour cropping up at different times, which seem to belong together by reason of similarities of the intonation, the vocabulary, syntax, the preoccupations in the utterance or to cohere as behaviour by reason of certain stereotyped gestures or mannerisms. It seemed therefore that one was in the presence of various fragments, or incomplete elements, of different 'personalities' in operation at the one time. Her 'word-salad' seemed to be the result of a number of quasi-autonomous partial systems striving to give expression to themselves out of the same mouth at the same time. ([Laing, 1960], 195-6)

Laing goes on to describe Julie's existence in ways that are eerily similar to the problems with autonomous agents we discussed in the last chapter — all the more eery because we are talking about actual, painful human experience and not a theoretical description of a machine: "Julie's being as a chronic schizophrenic was... characterized by lack of unity and by division into what might variously be called partial 'assemblies', complexes, partial systems, or 'internal objects'. Each of these partial systems had recognizable features and distinctive ways of its own" (197). Like the parts of behavior-based agents, each subsystem exists independently, with its own perception and action. Subsystems communicate, in Brooks' phraseology, 'through the world,' not by being integrated as a unified whole:

> Each partial system seemed to have within it its own focus or centre of awareness: it had its own very limited memory schemata and limited ways of structuring percepts; its own quasi-autonomous drives or component drives; its own tendency to preserve its autonomy, and special dangers which threatened its autonomy. She would refer to these diverse aspects as 'he', or 'she', or address them as 'you'. That is, instead of having a reflective awareness of those aspects of herself, 'she' would *perceive* the operation of a partial

> system as though it was not of 'her', but belonged outside. (198).[3]

While we can presume that artificial systems do not particularly care about being fragmented, for schizophrenic patients this feeling of coming apart, of losing life, of being reduced to a machine, is intensely painful. It is therefore ironic that psychiatric institutions themselves reinforce this feeling of mechanicity and lack of autonomous self. Erving Goffman, in his anthropological study *Asylums* [Goffman, 1961], analyzes psychiatric institutions, and concludes that one of their features is the attempted mechanization of their inmates.

Goffman's interest is in 'total institutions' such as psychiatric institutions, jails, and concentration camps, i.e. institutions that are barricaded from the rest of society and encompass all of their inmates' lives. From the beginning of an inmate's stay at such an institution, s/he is asked to give up his or her own identity in order to make for smoother processing by institutional bureaucracy.

> Admission procedures might better be called 'trimming' or 'programming' because in thus being squared away the new arrival allows himself to be shaped and coded into an object that can be fed into the administrative machinery of the establishment, to be worked on smoothly by routine operations. Many of these procedures depend upon attributes such as weight or fingerprints that the individual possesses merely because he is a member of the largest and most abstract of social categories, that of human being. Action taken on the basis of such attributes necessarily ignores most of his previous bases of self-identification. (16)

The admission procedures mark the beginning of a period of standardization, where inmates' individual identity is denied. "The admission procedure can be characterized as a leaving off and a taking on, with the midpoint marked by physical nakedness. Leaving off of course entails a dispossession of property, important because persons invest self feelings in their possessions. Perhaps the most significant of these possessions is not physical at all, one's full name; whatever one is thereafter called, loss of one's name can be a great curtailment of the self" (17). In place of patients' initial feeling of individuality, the institution enforces a homogeneous, standardized life, a homogeneity that is reflected in patients' environments, including their physical environment and clothing. "Once the inmate is stripped of his possessions, at least some replacements must be made by the establishment, but these take the form of standard issue, uniform in character and uniformly distributed. These substitute possessions are clearly marked as really belonging to the institution and in some cases are recalled at regular intervals to be, as it were, disinfected of identifications" (19).

The institutions' push to standardization, not only of patients' appearance, but of patients' very existence, is seen in the continuous intimate

[3] This splitting into subsystems is not the same thing as multiple personality. They are not experienced as completely separate individuals. In addition, Laing posits the subsystems as an explanatory mechanism that makes Julie's utterances more understandable; no one can directly know Julie's subjective experience, and she is not in a position to articulate it.

regulation of patients' lives: "[T]he inmate's life is penetrated by constant sanctioning interaction from above, especially during the initial period of stay before the inmate accepts the regulations unthinkingly.... The autonomy of the act itself is violated" (38). The nature of institutionalization is to (further) reduce patients' individuality and sense of autonomy. Patients must constantly ask for permission to do anything other than what the institution has planned for them; often times, even these requests are ignored, since patients may not be considered worth listening to. "Of course you had what they called an [sic] hearing but they didn't really want to hear you" ([Washington, 1991], 50). Over time, all resistance is worn down until patients passively accept the institution's decisions for them, becoming, at least in the eyes of its staff, little more than bureaucratic objects to be pushed and pulled into place.

Institutional Impoverishment of Meaning

So far, the mechanization of the inmate is similar in all total institutions. But psychiatric institutions are unique in that everything patients do — the last remaining bastion of individual expression — is treated as merely symptomatic. Patients are constantly monitored, their behavior continuously being examined for signs of illness.

> All of the patient's actions, feelings, and thoughts — past, present and predicted — are officially usable by the therapist in diagnosis and prescription.... None of a patient's business, then, is none of the psychiatrist's business; nothing ought to be held back from the psychiatrist as irrelevant to his job. (358)

In our everyday lives, we expect our utterances to be understood at face value; we become angry if, instead of trying to understand what we are saying, someone merely interprets it: "You are only so angry because you are still hypersensitive about your mother abandoning you." But in the institution, patients' words and actions are often simply interpreted as signs of illness. Rather than acting, patients *signify*. The patient's actions only function insofar as they are informational — they only *act* as ciphers, which it is then the responsibility and right of the doctor to decode. As a cipher, a patient's words can never be taken seriously as such; rather than being understood to refer to their intended meaning, the words are used to place the patient in the narrative of the doctor's diagnosis. "When you spoke, they judged your words as a delusion to confirm their concepts" ([Robear Jr., 1991], 19). The patient's acts are robbed of meaning so that another system of meaning can be imposed.

Maurice Blanchot expresses the frustrating abandonment of identity in this situation:

> I liked the doctors quite well, and I did not feel belittled by their doubts. The annoying thing was that their authority loomed larger by the hour. One is not aware of it, but these men are kings. Throwing open my rooms, they would say, 'Everything here belongs to us.' They would fall upon scraps of thought: 'This is ours.' They would challenge my story: 'Talk,' and my story would put itself at their service. In

> haste, I would rid myself of myself. I distributed my blood, my innermost being among them, lent them the universe, gave them the day. Right before their eyes, though they were not at all startled, I became a drop of water, a spot of ink. ([Blanchot, 1981], 14)

"Automatization eats away at things, at clothes, at furniture, at our wives, and at our fear of war." ([Shklovsky, 1990], 5)

The patient, rather than being treated as a full human being, is seen as a sign or symbol. Victor Shklovsky calls this reduction of a complex individual to a simple sign — a move which also occurs in AI when we reduce a complex behavior to a simple atom like 'hunting' or 'eating' — "automatization" [Shklovsky, 1990]. He argues that automatization causes one to forget the full richness of the actual object of automatization, replacing it with a single word. Similarly, by reducing patients to a set of signs to be interpreted, the institution only recognizes a small part of them.

The difficulty is that, once the bureaucratic system has standardized the patient, and the psychiatric system has ignored what the patient tries to say and do in favor of a symptomatic view, there is a huge gap between the the institution's mechanized view of the patient as symbol and the patient's experience of him- or herself as an individual person. The patient as a complete, subjective, and unique individual is simply not understandable under the rubric of the psychiatric institution. In this sense, the patient becomes invisible.

"I had been asked: Tell us '*just* exactly' what happened. A story?... I told them the whole story and they listened, it seems to me, with interest, at least in the beginning. But the end was a surprise to all of us. 'That was the beginning,' they said. 'Now get down to the facts.' How so? The story was over!" ([Blanchot, 1981], 18)

> The whole of me passed in full view before them, and when at last nothing was present but my perfect nothingness and there was nothing more to see, they ceased to see me too. Very irritated, they stood up and cried out, 'All right,where are you? Where are you hiding? Hiding is forbidden, it is an offense,' etc." ([Blanchot, 1981], 14)

The patient as understood by the institution is reified, atomized, mechanized, standardized, formalized, reduced to a mere ghost of his or her internally experienced self. Understood symptomatically, the patient's subjective experience is ignored. Susan Baur describes this limitation of the institutional approach to mental illness:

> I... believe that the medical model of mental illness excludes too much of the patient. Using this model, only parts of the patient are considered, and even when these parts are assembled by a multidisciplinary team into a manikin of a schizophrenic or of a manic-depressive, the spirit that animates the real person gets lost. Especially in chronic cases where mental illness and the desperately clever adaptations it inspires have become central to an individual's personality, the patient's own story and explanations — his delusions and imaginary worlds — must be included ([Baur, 1991], 105-6).

This leaves patients, sadly, unundderstood by the very institutions which are supposed to house and heal them.

Institutionalized Science

The fundamental problem of institutionalization is the bureaucratization of the patient. In bureaucracies, 'understanding' is reduced to categorization: instead of seeing each person as a unique, complex individual, people entering bureaucracies are classified and operated on according to standard, objective categories. Not many of those who read this thesis have been institutionalized; but all of us can recognize the feeling of frustration and alienation that comes from being treated as a thing by a bureaucrat.

The difficulty for institutional psychiatry is that, by treating the patient as a set of signs to be interpreted, the 'real' patient is left behind. The patient is formalized, reduced to a set of somewhat arbitrarily connected symptoms. Institutional psychiatry leaves the living patient out when it takes that formalized image of the patient for the patient him- or herself. The patient is no longer a living, unique, complex individual, but fragmented into a pile of signs: "she is autistic," "she shows signs of depersonalization," "she lacks affect."

This move — and the problems it brings — are paralleled in modern science. In science, the material, idiosyncratic properties of the objects to be studied are reduced to formal theories, preferably stated in terms of mathematics. While there is nothing wrong with formalization per se, difficulties come about when, as Katherine Hayles describes, the formalized theory is seen as more real than — or even *causing* — the material things being described [Hayles,]. One example of this is Dawkins's theory of the "selfish gene:" starting from theories of evolution, Dawkins argues that humans are 'really' no more than large bags of flesh whose only purpose is the propagation of genetic information [Dawkins, 1989] — thereby belittling the importance of the life history of individual living beings, which is only partly determined by genes. The same move is made in the institution: the patient is seen as fundamentally fragmented and symptomatic, structured as in psychiatric theory, not as a complex, embodied human being; his or her behavior is caused not by the patient's will but by a disease. According to Hayles, the information sciences sometimes make the same mistake: they see the world as 'really' a flow of information, with its materiality and noisy complexity an accidental after-effect.

Humanists will recognize reductionism.

In each of these areas, the wholistic and not-entirely-comprehensible aspects of the studied phenomena are forgotten, set aside in favor of a simpler and more elegant theory. But if your goal is to understand and engage in real life — or, in the case of AI, to be able to generate creatures that are in some sense truly alive — then it is best not to become too enamoured of your *theories* of life. If the only view of life you value is formalized and rationalized knowledge, then the world, which is probably neither formal nor rational, will always exceed it. The world is wholistic, complex, incompletely knowable; if only fragmented, elegant, and complete theories of the world are allowed, the actual world will seem to be incomprehensibly heterogeneous: schizophrenic. In this sense, schizophrenia in science is a result of the fragmentation that clean categorization brings about; it represents the limits of categorical knowledge.

Science and Alienation

So far, we have looked at science in terms of its relations to industrialization and to institutionalization. Industrialized science repeats the themes of assembly line work, using the processes of reification, specialization, atomization, standardization, formalization, and mechanization. Institutionalized science follows institutional psychiatry in reducing the objects of its study to formalized, fragmented versions of them. When both cases are taken to extremes, science loses something important: the subjective, idiosyncratic, incompletely knowable aspects of what it studies; the 'meaning' in life.

This is because industrialization and institutionalization share a negative side-effect: alienation. The term 'alienation' is used in multiple ways, but it can be fundamentally understood as a subjective feeling of being cut off: cut off from oneself, cut off from others, cut off from one's own actions. Under industrialization, workers are said to suffer from alienation because they are separated from the results of their work; instead of acting directly on their products, they merely tend machinery. Under institutionalization, the patient is alienated from the role which s/he is expected to play, and with which s/he may only marginally identify. Alienation is a fragmentation of life, a draining-away of the meaning of life, as the parts of one's life — one's self, one's friends, one's work — become separated, each part functioning atomically, with no subjective feeling of interconnection or wholism.

Modern science, too, is alienating. Unfortunately, the goal of reliable knowledge in science is often understood as necessitating a split between the individual scientist and the things or people which s/he studies, i.e. a subject / object divide. Science is generally understood not as a result of individual lives expressing themselves within a community of shared traditions, but as a self-contained, self-propelling force with its own logic, somehow only incidentally involving human beings. Even the very use of the word "I" in scientific papers is considered suspect; "the author is advised to avoid the use of first-person pronouns," as an otherwise extraordinarily helpful anonymous reviewer report of one of my papers elegantly circumlocuted it. "The experiment was conducted," "Results showed that... ," "It was noticed that... ," you read in the literature, as though research happened by itself, and the scientist only stopped by the office once a week to pick up the finished paper.

The scientist him- or herself is alienated in the sense that the product of his or her work — science itself — tends to feel independent of the scientist's personal existence. Indeed, the argument is often made, particularly in the natural sciences, that the individual scientist does not really matter; if a particular scientist had not done a certain piece of work, someone else would have done it. But in addition to the scientist itself, the very things that science studies are also alienated: they are atomized, fragmented, dissected, both literally and metaphorically; the very term 'science' probably comes from the Latin 'scindere,' to split [Gove, 1986].[4]

To be precise, it is unlikely that albino mice in a medical experiment have a subjective experience of alienation. But alienation can certainly

[4] Thanks to Stefan Helmreich for this observation.

be the experience of humans who try to understand themselves through the lens of science. Try being hospitalized for an unknown disorder, and watch the specialists turn you one by one into a skeletal system, a gastro-intestinal system, a nervous system, an immune system, and, if it turns out your problem is wholistic, a hysteric. Try to understand what makes you tick by reading the latest results in experimental psychology and statistical sociology; the more studies you read, the more multiple you feel, the less you are able to synthesize them into a coherent worldview. Seen through the lens of science, you are split into biology, psychology, and sociology, and in each of these realms into a thousand more subfields and experimental results. Good luck finding yourself!

Alienation is bad for science because it makes the things science studies seem fragmented. Science breaks things into pieces to study them; whether or not they 'actually' are fragmented (probably not), they end up looking that way to us. This means that the results of science can be misleading. In this section, we will look at several ways of doing science that try to resolve the problems of objectivist science. First, we will look again at schizophrenia — now understood in psychiatric terms — to understand concretely in this example how objectivist science, in alienating doctor from patient, can unconsciously fragment the patient, rendering him or her unnecessarily incomprehensible. We will then look at an alternative approach proposed by anti-psychiatry to find alternative ways of doing science that avoid the pitfalls of alienation.

Alienation and Objectivist Science: The Divided Self

Earlier, we noted that schizophrenia sometimes includes an alienation-from-self in that the self is experienced as split into different parts. R. D. Laing describes schizophrenia as including, not just a division within the parts of the self, but also a disruption between the self and the rest of the world.

> The term schizoid refers to an individual the totality of whose experience is split in two main ways: in the first place, there is a rent in his relation with his world and, in the second, there is a disruption of his relation with himself. Such a person is not able to experience himself 'together with' others or 'at home in' the world, but, on the contrary, he experiences himself in despairing aloneness and isolation; moreover, he does not experience himself as a complete person but rather as 'split' in various ways, perhaps as a mind more or less tenuously linked to a body, as two or more selves, and so on. ([Laing, 1960], 17)

"The standard texts contain the descriptions of the behaviour of people in a behavioural field that includes the psychiatrist. The behaviour of the patient is to some extent a function of the behaviour of the psychiatrist in the same behavioural field. The standard psychiatric patient is a function of the standard psychiatrist, and of the standard mental hospital. The figured base, as it were, which underscores all Bleuler's great description of schizophrenics is his remark that when all is said and done they were stranger to him than the birds in his garden." ([Laing, 1960], 28)

Laing describes how schizophrenics may construct a 'false-self' system, through which they present a false front to the world, while keeping their self-identified 'real' selves safely hidden away. This false-self mechanism may then be partly responsible for a patient's further deterioration; without the confirmation of self that social interaction brings, patients' real selves are in danger of wasting away.

This split between a schizophrenic and their surrounding environment has been more generally noted. Schizophrenic language itself may lack reference to context; a patient may, for example, be laughing while recounting a heart-rending story [American Psychiatric Association, 1994]. Social withdrawal or a 'break with reality' is also common. "The person may be so withdrawn from the world that h/she is absorbed entirely in his/her mixed-up thoughts" ([Webb *et al.*, 1981], 72).

Understanding schizophrenics can therefore be difficult because it is often hard to establish social contact with them. Because schizophrenics may fear true social contact, they may even actively undermine the doctor's understanding. "A good deal of schizophrenia is simply nonsense, red-herring speech, prolonged filibustering to throw dangerous people off the scent, to create boredom and futility in others. The schizophrenic is often making a fool of himself and the doctor" ([Laing, 1960], 164). This complicated the doctor's job; it is simply hard to understand someone who refuses to interact.

For psychiatrists like Laing, one of the main avenues toward understanding schizophrenia, then, is to break down the barrier between schizophrenic patients and their social worlds by engaging in personal relationships with them, i.e. by putting patients back into their social contexts. But Laing finds that the methods and language of clinical psychiatry actually undermine his goal to connect with the patient as a human being. This is because, rather than treating the patient as a person, psychiatrists see patients as a bundle of symptomatology. Mechanistic explanations reduce the patient to a bundle of pathological processes.

This 'clinical detachment,' by which the patient can be seen as a mere instance of a disease, is considered good because treating the person as a whole person would mean entering into a personal relationship with them, undermining objectivity.

> [T]here is a common illusion that one somehow increases one's understanding of a person if one can translate a personal understanding of him into the impersonal terms of a sequence or system of *it*-processes. Even in the absence of theoretical justifications, there remains a tendency to translate our personal experience of the other as a person into an account of him that is depersonalized. (22)

But just as it is inaccurate to describe an animal or object in anthropomorphic terms, it is equally inaccurate to picture a human as an animal or automaton.

"As a psychiatrist, I run into a major difficulty at the outset: how can I go straight to the patients if the psychiatric words at my disposal keep the patient at a distance from me? How can one demonstrate the general human relevance and significance of the patient's condition if the words one has to use are specifically designed to isolate and circumscribe the meaning of the patient's life to a particular clinical entity?" [Laing, 1960], 17)

Fundamentally, the stumbling block for objectivist psychiatry is that a detached, impersonal attitude does not lead to a view of the patient independent of the psychiatrist's personal attitudes. This is because the objective, clinical approach that psychiatrists may take is itself part of the schizophrenic patient's situation. The 'objectivity' the psychiatrist takes on itself influences what the patient does and how the psychiatrist can come to understand him or her.

> The clinical psychiatrist, wishing to be more 'scientific' or 'objective', may propose to confine himself to the 'objectively' observable behaviour of the patient before him. The simplest reply to this is that it is impossible. To see 'signs' of 'disease' is not to see neutrally.... We cannot help but see the person in one way or other and place our constructions

> or interpretations on 'his' behaviour, as soon as we are in a relationship with him. (31)

Even the objectivist psychiatrist is constructing a particular kind of relationship with the patient, one that cuts off the possibility of human understanding. By treating the patient as separate, as not a person, as a thing, the patient as human is rendered incomprehensible.

Laing argues that institutional psychiatric practice cannot fully understand schizophrenia because it actually *mimics* schizophrenic ways of thinking, depersonalizing and fragmenting patients. "The most serious objection to the technical vocabulary currently used to describe psychiatric patients is that it consists of words which split man up verbally in a way which is analogous to the existential splits we have to describe here" (19). Clinical language atomizes and reifies patients, studying them in isolation from their worlds and from the psychiatrist.

> Unless we begin with the concept of man in relation to other men and from the beginning 'in' a world, and unless we realize that man does not exist without 'his' world nor can his world exist without him, we are condemned to start our study of schizoid and schizophrenic people with a verbal and conceptual splitting that matches the split up of the totality of the schizoid being-in-the-world. Moreover, the secondary verbal and conceptual task of reintegrating the various bits and pieces will parallel the despairing efforts of the schizophrenic to put his disintegrated self and world together again. (19-20)

By studying schizophrenics in isolation and in parts, psychiatry threatens to itself become schizophrenic, and schizophrenics incomprehensible.

Anti-Psychiatry: Science in Context

If objectivist psychiatry distorts and fragments schizophrenia, rendering it incomprehensible, are there other ways of doing science that avoid alienation? Laing and other sympathetic colleagues in the 60's and 70's, termed *anti-psychiatrists* for their opposition to mainstream psychiatry, suggest that the schizophrenizing aspects of institutional psychiatry can be avoided by understanding schizophrenia in the context of the patient's life. If schizophrenia is to be understood, anti-psychiatrists argue, we need to think of schizophrenics, not as self-contained clusters of symptoms, but as complex humans. This means studying them, not in a vacuum, but in relation to both their lifeworlds and to the people who study and treat them, including psychiatrists themselves.

"A schizophrenic out for a walk is a better model than a neurotic lying on the analyst's couch. A breath of fresh air, a relationship with the outside world." ([Deleuze and Guattari, 1977], 2)

The difference between these approaches can be understood by contrasting objectivist and subjectivist descriptions of patient behavior. The clinical approach reifies the patient's behavior into a cluster of pathological symptoms, with no apparent relation to each other or the patient's broader life experience:

> [S]he had auditory hallucinations and was depersonalized; showed signs of catatonia; exhibited affective impoverishment and autistic withdrawal. Occasionally she was held to be 'impulsive.' ([Laing and Esterson, 1970], 32)

The phenomenological approach advocated by anti-psychiatrists, on the other hand, tries to understand the patient's experience of herself as a person:

> [S]he experienced herself as a machine, rather than as a person: she lacked a sense of her motives, agency and intentions belonging together: she was very confused about her autonomous identity. She felt it necessary to move and speak with studious and scrupulous correctness. She sometimes felt that her thoughts were controlled by others, and she said that not she but her 'voices' often did her thinking.

Anti-psychiatrists believe that statistics and symptomatology, the foundations of institutional psychiatry, are misleading because they reduce the patient to a mass of unrelated signs. Instead of leading to a greater understanding of the patient, the patient's subjective experiences are lost under a pile of unconnected data.

> It is just possible to have a thorough knowledge of what has been discovered about the hereditary or familial incidence of manic-depressive psychosis or schizophrenia, to have a facility in recognizing schizoid 'ego distortion' and schizophrenic ego defects, plus the various 'disorders' of thought, memory, perceptions, etc., to know, in fact, just about everything that can be known about the psychopathology of schizophrenia or of schizophrenia as a disease without being able to understand one single schizophrenic. Such data are all ways of *not* understanding him. ([Laing, 1960], 33)

Instead of trying to extract objectively verifiable data about the patient, anti-psychiatrists believe psychiatry should be based on *hermeneutics*, a subjective process of interpretation which aims for a better understanding of the way in which the schizophrenic patient experiences life. Laing finds that when schizophrenic patients are treated 'subjectively' — that is to say, when attempts are made, not to catalog their symptoms, but to understand their phenomenological viewpoints, even when they include such apparently alien components as delusions or hallucinations — schizophrenia can be made much more comprehensible. In *Sanity, Madness, and the Family*, Laing and Esterson give 11 case studies of schizophrenic patients whose behavior, initially incomprehensible and even frightening, is made understandable by putting it in the context of the patient's family life. For example, a patient with a delusion that other people are controlling her thoughts is found to live in a family where her parents undermine every expression of independent thought, telling her that they know better than her what she thinks.

It is important to note that understanding a schizophrenic patient is not the same as curing him or her. Giving meaning to delusions and hallucinations does not take them away or reduce their effect on a patient's life. Nevertheless, complementing clinical understanding of a patient with phenomenological interpretation of the patient's life-world gives a fuller picture of the patient as human being and provides better understanding of the nature of schizophrenia in this individual person.

Anti-psychiatrists believe that the concept of schizophrenia as a pathological disorder affecting individuals in isolation is misleading. When studied in context, schizophrenic symptomatology that otherwise seems bizarre and inexplicable starts to make sense; in this sense, schizophrenia is a sane response to an insane situation. Anti-psychiatrists note that schizophrenic patients are sometimes the locus of negative tension in their families; they hypothesize that patients may take on the label of 'sick' so that their families can avoid introspection into the negative aspects of their psychodynamics. In addition, cultural influences — the broader atomization and depersonalization of post-industrial life — may itself be 'schizophrenizing,' a factor which is forgotten when research focuses on the sole, sick individual instead of the society that in some sense causes his or her illness. Finally, the very reification of schizophrenia as a disease an individual 'has' is misleading, because it separates a patient from his or her behavior and pathologizes it.

In essence, anti-psychiatrists make not only an epistemological argument, but an ethical one. According to anti-psychiatrists, the use of schizophrenia in institutional psychiatry is not only incorrect, but morally wrong. Treating people as objects not only leaves them incomprehensible in their humanity; it also makes it easier to treat them as objects, cogs in the institutional machine. Depersonalization is not only an intellectual viewpoint, but the daily experience of institutionalized patients, which ranges from mild annoyances in exclusive, private wards, to the warehousing of humanity in large, public institutions, to the absolutely inhuman conditions of institutes for the criminally insane (see e.g. [Viscott, 1972]). "We are a special breed of farmyard animals," as Sylvain Lecocq wrote his doctor a year and a half before hanging himself from his hospital bed ([Lecocq, 1991], 160).

Anti-psychiatrists often antagonize more mainstream psychiatrists, in much the same way that the cultural critics of science antagonize scientists. In essence, the anti-psychiatrists argue that schizophrenia is a social construct, supported by the medical and institutional establishments, but not necessarily particularly helpful in treating those considered mentally ill. Psychiatrists interpret this as an argument that schizophrenia is a fiction, a mere social label, and that objectivist psychiatry is, in essence, colluding with families to label otherwise perfectly healthy people as dysfunctional. And some anti-psychiatrists basically agree with this perception:

> [S]chizophrenia is a micro-social crisis situation in which the acts and experiences of a certain person are invalidated by others for certain intelligible cultural and micro-cultural (usually familial) reasons, to the point where he is elected and identified as being 'mentally ill' in a certain way, and is then confirmed (by a specifiable but highly arbitrary labelling process) in the identity 'schizophrenic patient' by medical or quasi-medical agents. ([Cooper, 1967], 2, emphasized in original)

One of the results of this mutual antagonism is a backlash in institutional psychiatry, as psychiatrists attempt to disprove the unattractive claims of anti-psychiatry by showing that schizophrenia is basically a biological illness which can be objectively identified. Anti-psychiatry was dealt another blow in the 80's, when its demonization of institutionalization was used as a pretext for the economically attractive Reagan-era depopulation and closure of mental hospitals. The former inmates, now dumped on the streets basically untreated and unable to cope with life, can be seen in most major American urban centers, an apparent living testament to anti-psychiatry's bankruptcy — if one ignores the fact that anti-psychiatrists never proposed getting rid of the problems of institutions by simply kicking all the patients out.

But even the *Diagnostic and Statistical Manual of Mental Disorders*, which represents the conservative mainstream of psychiatry, notes the sociocultural face of schizophrenia: that, for example, schizophrenia is more prevalent and harder to treat in industrialized nations [American Psychiatric Association, 1994]. Schizophrenia is probably not *merely* a social label in the way more extreme anti-psychiatrists seem to imply — it is, for example, more prevalent among relatives of already-diagnosed schizophrenics, even when raised apart. But, at the same time, schizophrenia is also clearly influenced by environmental factors (for example, it is not unusual for only one half of a monozygotic twin to have it). Schizophrenia clearly *does* depend on the sociocultural context within which the labeled schizophrenic lives. The anti-psychiatric interest in contextualization therefore lives on, even in mainstream psychiatry.

Alternatives to Alienated Science

Anti-psychiatry rejects the objectivist stand of institutional psychiatry, arguing that understanding human beings is qualitatively different from understanding inanimate objects as in physics.

> It may be maintained that one cannot be scientific without retaining one's 'objectivity.' A genuine science of personal existence must attempt to be as unbiased as possible. Physics and the other sciences of things must accord the science of persons the right to be unbiased in a way that is true to its own field of study. If it is held that to be unbiased one should be 'objective' in the sense of depersonalizing the person who is the 'object' of our study, any temptation to do this under the impression that one is thereby being scientific must be rigorously resisted. Depersonalization in a theory that is intended to be a theory of persons is as false as schizoid depersonalization of others and is no less ultimately an intentional act. Although conducted in the name of science, such reification yields false 'knowledge'. It is just as pathetic a fallacy as the false personalization of things. ([Laing, 1960], 24)

The belief in objectivity — in the sense of belief that the psychiatrist as a knowing subject can be cleanly divided from the patient, who is an object to be understood mechanically — fundamentally distorts our perception of patients, simply because patients are always already in a human relationship with the doctor, even when that relationship consists of the doctor ignoring the patient's humanity.

Anti-psychiatrists not only criticize objectivist science as alienated and alienating; they also develop new ways of achieving the goals of psychiatry that do not have the same flaws. Anti-psychiatry argues that symptomatic views of mental patients actually reinforce schizophrenia by depersonalizing patients, fragmenting them, and removing them both physically and epistemologically from their contexts. Instead, anti-psychiatrists develop a new practice, one that is based on respect for the patient as a complete person and attempts to interpret his or her behavior not in isolation but in the context of the patient's complete lifeworld.

This context is not just limited to the patient's family. The really novel step the anti-psychiatrists take is to become aware of the role *they themselves* play in interpreting and interacting with the patient. Anti-psychiatrists do not see themselves as looking on the patient's life from the outside; they understand that even as they are trying to study the patient in as unbiased a way as possible, they cannot help but be in a human relationship with the schizophrenic that effects how they come to understand the patient him- or herself.

The fundamental recommendation anti-psychiatry makes for the methodology of psychiatry is this: *the patient should be studied in context.* This means on the one hand that the 'parts' of the patient — his or her symptoms, 'subsystems,' actions, language – should be studied in relation to one another, forming a unified rather than fragmentary picture of the patient as a person. On the other hand, it means that the patient should be studied in a social context, a context which includes the people who are judging him or her.

This proposal for addressing the problems of alienated science is similar to ones that have been raised in other fields. In neurology, for example, Kurt Goldstein argues that the fragmentation of organisms as necessarily occurs in science is insufficient for understanding them, since in life they function wholistically [Goldstein, 1995].

"To attempt to understand life from the point of view of the natural-science method alone is fruitless." ([Goldstein, 1995], 18)

> We have said that life confronts us in living organisms. But as soon as we attempt to grasp them scientifically, we must take them apart, and this taking apart nets us a multitude of isolated facts that offer no direct clue to that which we experience directly in the living organism. Yet we have no way of making the nature and behavior of an organism scientifically intelligible other than by construction out of facts obtained in this way. (27)

Goldstein argues that in order to understand complete organisms, one needs to balance fragmenting and symptomatizing methods from science with a more humanistic interest in how individuals function as a whole within the context of their lives. "Certainly, isolated data acquired by the dissecting method of natural science could not be neglected if we were to maintain a scientific basis. But we had to discover how to evaluate our observations in their significance for the total organism's functioning and thereby to understand the structure and existence of the individual person" (18). In practice, this means that Goldstein does not simply look for signs and symptoms, but tries to understand symptoms as fragmented manifestations of wholistic alterations to an individual nervous system that occur under disease. Statistics, he argues, is useless for this kind of understanding; instead, Goldstein works with case studies, "in which the historical, the personal, the experimental, and the clinical could all be brought together as a unity" ([Sacks, 1995], 8).

From Alienation to Wholism

The common thread in these solutions to the fragmentation of science is to combat alienation — the separation of scientist from object of science, the separation of different scientific subfields, the separation of the parts of the object being studied — by adding wholism to the toolbox of science. The object itself should be studied as a whole, its 'parts' being understood in relation to one another and to the object or organism as a whole. In addition, the object itself should be understood in context, a life world which includes the scientists studying the object. Rather than cutting the scientist off, the scientist and the object should be understood as in relationship to one another, leading to a 'personal' or 'subjective' science. In a wholistically informed science, 'objectivity' — in the sense of a natural world to be studied independently of the people who study it — is not possible; instead of objectivity, the goal of subjectivist science is, as Varela et. al. put it, *disciplined* knowledge [Varela *et al.*, 1991].

The alert reader may recognize the postulates of anti-boxology as expressed in Chapter 1. In essence, this chapter has been an articulation of the reasons for the anti-boxological approach. In the next chapter, we will look at the implications of this approach to science for AI, and in particular agent design. I will argue that autonomous agents, like schizophrenic patients, are cut off from their context; like assembly line workers, they are split into parts and rationalized until their overall actions lose any meaning. The result of these two moves is schizophrenia. To combat them, we can rethink AI's methodological strategies by importing the contextualizing approach of anti-psychiatrists and other critics of objectivist science. I call the resulting approach to AI "socially situated AI," and use it as the basis for rethinking agent design in the rest of the thesis. But first, we will take a short break to look at the system, the Industrial Graveyard, that both demonstrates the concepts of the analysis in this Chapter and provides the testbed for the technical work of the thesis.

Intermezzo I

The Industrial Graveyard

One of the heuristics we can derive from the previous chapter is that agents should be studied in the context in which they are used. In this Intermezzo, I will explain the system which provides the context for the agents developed for this thesis. This system, the *Industrial Graveyard*, is intended to demonstrate both the technical and the theoretical ideas of this thesis.

The Industrial Graveyard is a virtual environment in which a discarded lamp (the Patient, top right) ekes out a marginal existence in a junkyard. It is overseen by a nurse/guard (the Overseer, bottom right) from the Acme Sanitation and Healthcare Maintenance Organization. In this scenario, users are asked to take on the role of an auditor overseeing the efficiency of the Acme-run junkyard. Their job is to make sure the Overseer is sufficiently interceding in the Patient's existence. Here, I will describe the Industrial Graveyard both technically and in its connections to the theoretical ideas of the last chapter.

Introduction

The Industrial Graveyard is intended to make the user feel viscerally the constraints of objectivist knowledge production. There are two levels at which these constraints work. First of all, the Patient, for whom users generally develop a sense of pity, is shown caught within an industrial-institutional nightmare. The Patient has been discarded and lives in a fenced-in junkyard, in which its only companion is an Overseer who constantly punishes the Patient. The Patient is judged objectively by the Overseer, which is to say, without personal consideration of the meaning of the Patient's actions. When the Patient is no longer efficiently manageable, it is killed.

The second level at which the constraints of objectivist knowledge production are demonstrated is at the meta-level of the technology itself. In the rhetoric of virtual environments, you can do anything — be anyone — in a virtual world that lacks the limitations of everyday, physical existence. But technical systems always contain both consciously and unconsciously imposed constraints on what users can do, whether from the limitations of input technology (e.g., you can only take actions which correspond to a simple verb in the system's vocabulary) or simply because

ACME Sanitation & HMO

Welcome to our organization! We are proud to be the nation's largest and fastest growing Sanitation and Healthcare Maintenance Organization. We believe patients are best served through fast, efficient service. We strive to give patients what they need while minimizing costs through the reduction of extraneous services. You can be proud you have chosen to become part of a well-functioning health maintenance and sanitation machine.

Your input profile has indicated your appropriateness for supervisory position #45-TBKJ. This document contains instructions for your role.

You have been assigned to the Sanitation and Disposal sector. When patients can no longer function properly in their societal role, they can become a burden to themselves and those around them. Acme S&D is proud to take on the responsibility of their care. At the same time, in order to maintain profitability, dysfunctional patients must be monitored particularly closely, since they suffer from a chronic condition and as such may incur high costs over the lifetime of the patient. Patients are therefore assigned automated overseers which monitor their behavior. These overseers provide necessary care, but lack the human intuition to always determine when patient behavior is malignant. Your job is to provide back-up for the overseer, ordering it, when necessary, to monitor the patient more closely.

FIGURE I.1: User's introduction to the Industrial Graveyard

the authors did not think to program in some option that users can think of (have you ever tried to make friends with the monsters in Doom?). Because virtual environments are often presented in current rhetoric as authorless — as real worlds, not personal visions — they, too, are a form of objectivist knowledge production.

In this sense, the Industrial Graveyard can be understood as a *parody* of a virtual environment. The function of parody in the system is to make objectivist construction of technical artifacts, which is normally a theoretical construct, be experienced in a visceral sense by users, becoming part of their subjective experience. This is done by exaggerating the constraints of the system to the point where users are forced to become aware of them. Far from being able to be anyone or do anything, users are told exactly what they are expected to do, and the system is designed to try to make them uncomfortable in the role to which they are assigned. The 'cartoony' nature of the world, in contrast with the photographic physical realism of many virtual environments, is also intended to communicate that the world was written by someone.

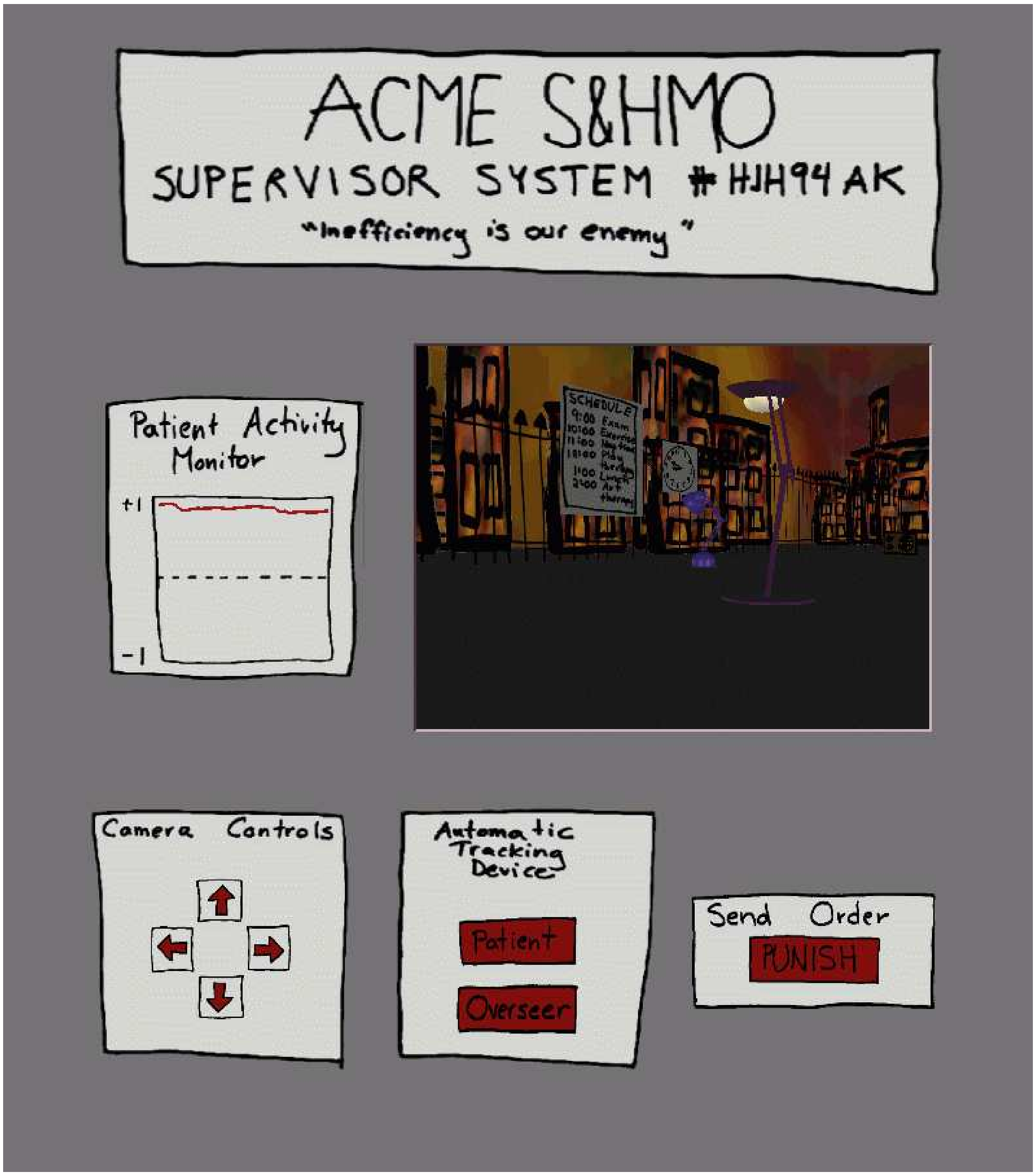

FIGURE I.2: The presentation of the system.

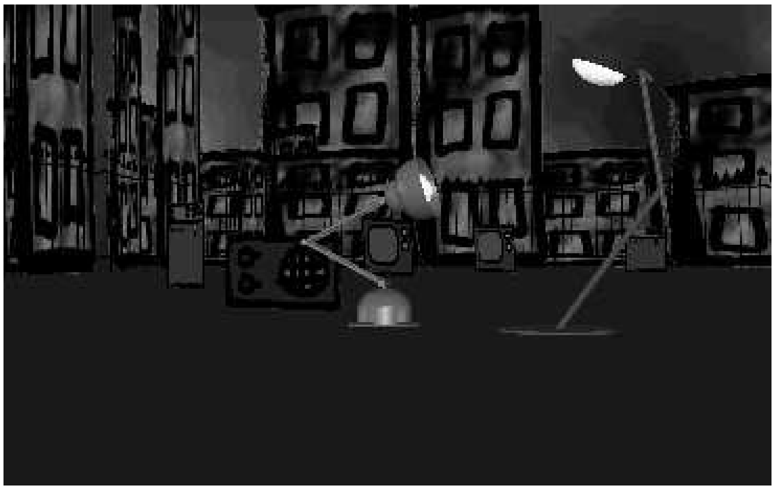

FIGURE I.3: The Overseer prepares to attack the Patient.

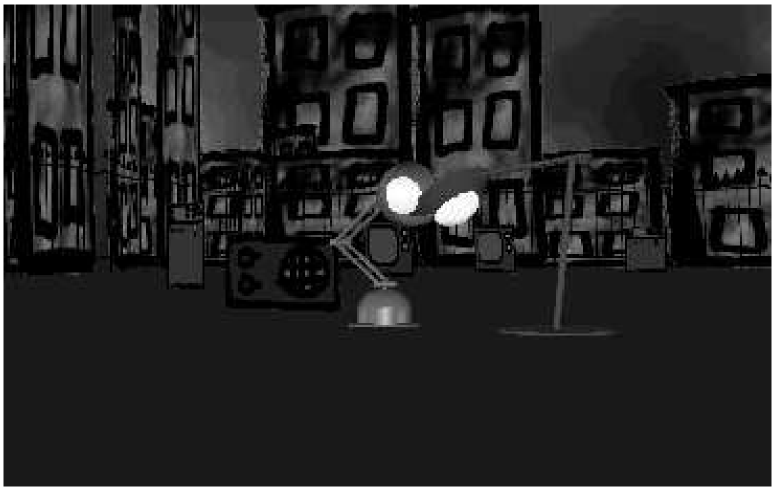

FIGURE I.4: The Overseer strikes the Patient.

Design

Users are introduced to the system through a set of written instructions that explain their role (Figure I.1). The instructions repeat the themes of industrialized, mechanized culture, and place users in a position of colluding with the Overseer against the patient.

Users then 'interact' with the system, which is illustrated in Figure I.2. The word 'interact' is in quotation marks because users' ability to influence the system is minimal. They can move the "surveillance camera" around (although it stays within a fenced-off area), and they can order the Overseer to harass the Patient. There is nothing users can do to help the Patient.

To the left of the view into the junkyard is a graph which shows the user how good or bad the Patient is being. 'Goodness' is calculated objectively by measuring the amount of movement of the angles of the

FIGURE I.5: The aftermath of the attack.

Patient's body. The Patient is considered optimally 'good' when it is frozen in place. When the Patient becomes 'bad' — by, for example, being excited about exploring the junkyard — the Overseer comes over and strikes the Patient (Figures I.3-I.4), turning the Patient off (Figure I.5).

Plot

The Industrial Graveyard includes a kind of story — the Patient is deposited in the junkyard, explores it (under constant interruptions by the Overseer), and, eventually, is killed by the Overseer. The 'plot' of the Industrial Graveyard follows Tinsley Galyean's notion of interactive narrative flow: it accommodates users' actions and random variations in the agents' behavior without fundamentally altering the story [Galyean, 1995]. Variations occur in the timing of the plot points and in how they are realized, but, no matter what, the same basic plot points always occur.

The story is maintained using the concept of "story stages," which are component pieces of the story in a sequential order. The current stage is stored in a data structure which is accessible to both characters. When a character does something to advance the story to the next stage, the character updates the data structure to reflect the new story stage. Both characters modify their behavior according to the stage the story is in.

To start out with, the Patient is dropped into the world, landing in a diagnostic machine. The Overseer comes over and reads the Patient's diagnosis, while the Patient cowers (Figure I.6). After the Overseer leaves, the Patient looks around, and gingerly steps out into the junkyard.

The Patient wanders around the junkyard, looking at the objects in it, and trying to stay away from the Overseer, who regularly harasses it. Eventually, the Patient notices a schedule of activities posted on the fence. It becomes engrossed in reading it (Figure I.7), oblivious to the Overseer, who comes up behind it. The clock turns 10 (time to exercise, according to the schedule), and the Patient, noticing the Overseer, frantically starts exercising.

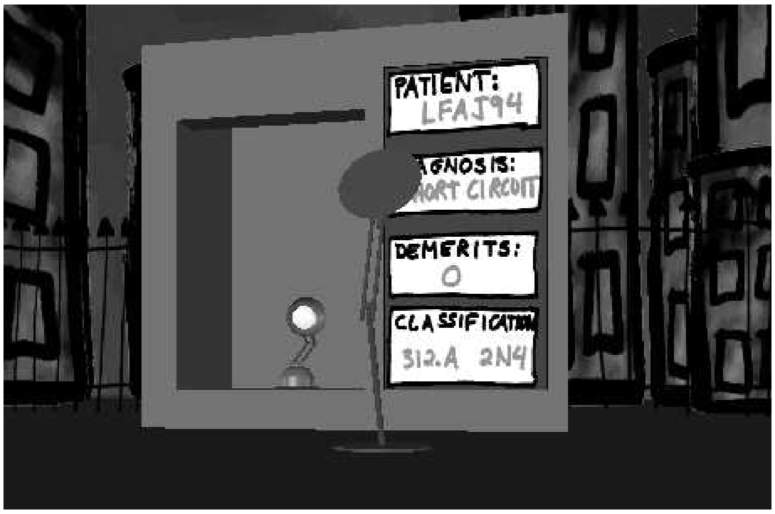

FIGURE I.6: The Patient is examined in the monitor.

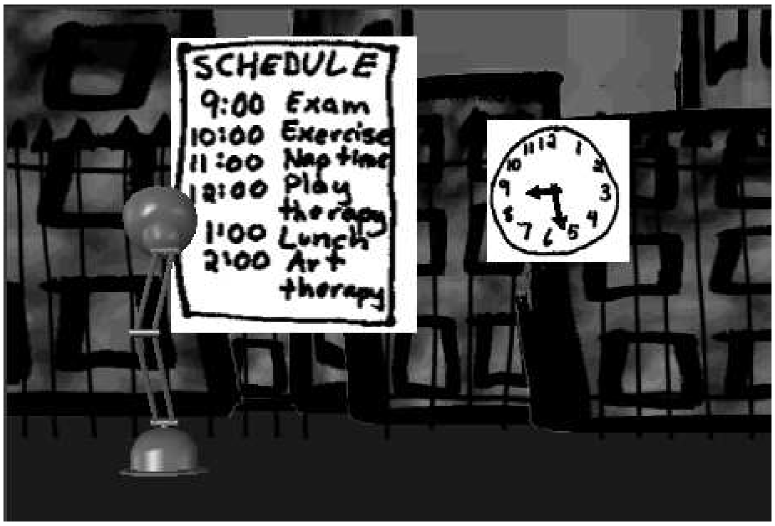

FIGURE I.7: The Patient reads the schedule.

After the Overseer leaves, the Patient loses interest in exercise and wanders off sadly. It stands by the fence, sighing and looking out at the world that has rejected it. Suddenly, the Patient's light goes out. The Patient shakes its head, trying to get the light on. When that doesn't work, the Patient starts hitting its head on the ground, trying to fix the short circuit (Figure I.8). It gets more and more frantic, banging around more and more — and therefore, by the logic of the Overseer, being more and more bad.

Finally, the Overseer comes over. The Patient cowers, wondering what is going on. The Overseer brings a large mechanism over the Patient's head, from which a beam emerges (Figure I.9). When the beam recedes, only the Patient's corpse is left (Figure I.10).

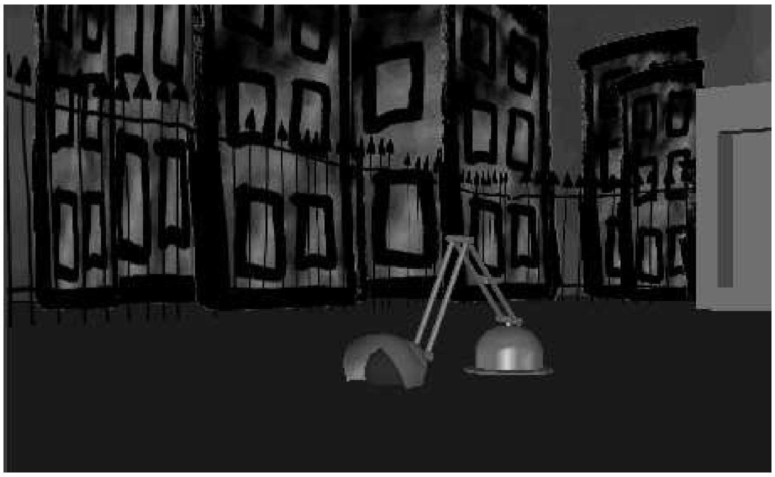

FIGURE I.8: The Patient hits its head on the ground.

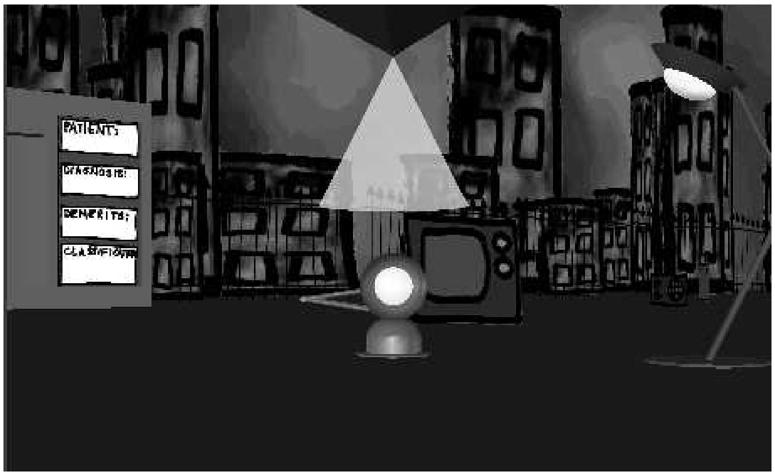

FIGURE I.9: The Patient being struck by the beam.

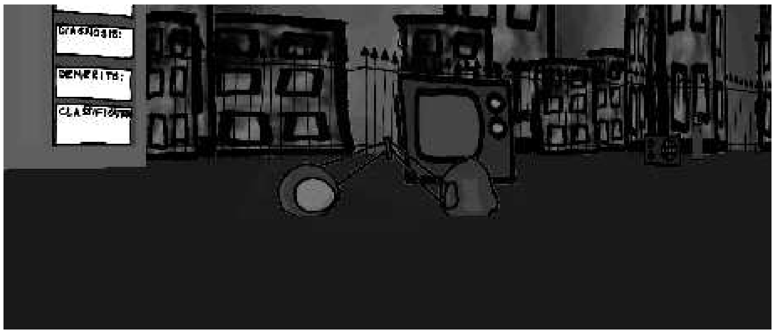

FIGURE I.10: The happy ending.

Construction

The Industrial Graveyard is built on the skeleton of a previous Oz group system, the *Edge of Intention* [Loyall and Bates, 1993]. In order to create the Industrial Graveyard, I removed the Edge of Intention's hand-coded graphics system (bouncing balls in a 2.5-D environment with fixed camera position) and replaced them with an interface to the Inventor 3D graphics toolkit [Wernecke, 1994]. The agent's bodies and world are Inventor models which can be loaded and reconfigured at run-time; the user's viewing position is a movable camera immersed in the world, rather than a God-like view from above. The 'cartoony' flat objects in the Industrial Graveyard are created by projecting transparent texture maps onto flat planes.

The Overseer's behavior is written in Bryan Loyall's Hap [Loyall and Bates, 1991], while the Patient is written in the *Expressivator*, the system I will describe in Chapters 5 and 7. Each agent architecture sends low-level commands ("spin," "jump," "move eyes") to a motor system which models the creature's bodies. An underlying physical simulation implements actions by modelling the lamps as Edge-of-Intention-style bouncing spheres. The system runs in real time on an SGI Indigo 2. Most of the running time is devoted to graphics; the Patient's mind takes about 14 milliseconds per frame, while the graphics takes about 77. Most of the graphics time is devoted to texture mapping.

Chapter 4

Socially Situated AI

The cultural theory analysis of Chapter 3 suggests that industrialization and institutionalization share properties that lead to schizophrenia. Both industrialization and institutionalization take objective views of living beings. By 'objective,' I mean that they are taken out of their sociocultural context and reduced to a set of data.[1] Because these data are not related to one another or the context from which they sprung, the result is a fragmentation of experience that cultural theorists term schizophrenia.

Cultural theory therefore suggests that, in order to address schizophrenia, we can take the *opposite* approach. Rather than seeing workers or patients as objects to be manipulated or diagnosed, we could see them *subjectively*. This means turning objectivity as defined above on its head: studying people in their life context and relating the things we notice about them to their existence as a whole.

If you are a technical researcher, it is quite possible that Chapter 3 was an insurmountable struggle, or at the very least left you with lingering doubts about the accuracy or validity of the cultural theory argument. But however you feel about the understandability or truth-value of that argument, the perspective cultural theory brings can be understood as a kind of heuristic which could be tried out in AI. At this level, cultural theory suggests the following: *if your agents are schizophrenic, perhaps you need to put them in their sociocultural context*.

In this chapter, we'll explore what it means for an agent to be designed and built with respect to a sociocultural environment. This way of doing AI I term *socially situated AI*. I will differentiate socially situated AI from the approaches taken in classical and alternative AI, and then discuss the impact this methodological framework has on the way AI problems are defined and understood. This different way of doing AI will become the key to solving schizophrenia in Chapters 5 and 7 by suggesting the redefinition of the problem of schizophrenia as a difficulty of *agent communication* rather than of *internal agent structure* — thereby finding a trapdoor to get us out of the Catch-22 of schizophrenia and atomization.

[1]The notion of what exactly objectivity means in various fields and usages is a quagmire in which, at the moment, I prefer not to be morassed. Please accept this usage of objectivity as a definitional statement of what I mean by 'objectivity' here, as opposed to a pronouncement of what anyone would mean by it.

AI in Context

The heuristic suggested by cultural theory — that agents should be considered with respect to their context — should have a familiar ring to technical researchers. The contextualization of agents, i.e. their definition and design with respect to their environment is, after all, one of the major bones alternativists like to pick with classicists. Alternative AI argues that agents can or should only be understood with respect to the environment in which they operate. The complexity or 'intelligence' of behavior is said to be a function of an agent *within* a particular environment, not the agent understood in isolation as a brain-in-a-box.[2]

But the contextualization which is so promoted in alternative AI is actually limited, in particular by the following implicit caveat to its methodology: *the agent is generally understood purely in terms of its* physical *environment* — *not* in terms of the sociocultural environment in which it is embedded. Generally speaking, alternativists examine the dynamics of the agent's activity with respect to the objects with which the agent interacts, the forces placed upon it, and the opportunities its physical locale affords. Some alternativists have also done interesting work examining the dynamics of agent activity in *social* environments, where 'social' is defined as interaction with other agents. They generally do not, however, consider the *sociocultural* aspects of that environment: the unconscious background of metaphors upon which researchers draw in order to try to understand agents, the social structures of funding and prestige that encourage particular avenues of agent construction, the cultural expectations that users — as well as scientific peers — maintain about intentional beings and that influence the way in which the agent comes to be used and judged.

In fact, when such aspects of the agent's environment are considered at all, many alternativists abandon their previous championing of contextualization. They see these not-so-quantifiable aspects of agent existence not as part-and-parcel of what it means to be an agent in the world, but as mere sources of noise or confusion that obscure the actual agent. They may say things like this: "The term 'agent' is, of course, a favourite of the folk psychological ontology. It consequently carries with it notions of intentionality and purposefulness that we wish to avoid. Here we use the term divested of such associated baggage" ([Smithers, 1992], 33) — as though the social and cultural environment of the agent, unlike its physical environment, is simply so much baggage to be discarded.

In this respect, the alternativist view of agents-in-context is not so different from the Taylorist view of worker-in-context or the institutional view of patient-in-context. After all, Taylorists certainly look at human workers in context; in the terminology of situated action, they analyze and optimize the ongoing dynamics of worker-and-equipment within the situation of a concrete task, rather than the action of the worker alone and in general. Similarly, institutional psychiatrists look at human patients in context; they are happy to observe and analyze the dynamics of patient interaction with other people and objects in the world, as long as in those observations and analyses they do not need to include themselves. In each of these cases, contextualization is stopping at the same point: where the

[2] Classicists will recognize the same argument as Simon's ant.

social dynamics between the expert and the object of expertise, as well as its *cultural* foundation, would be examined.

I do not believe that the elision of sociocultural aspects from the environment as understood by alternative AI is due to any nefarious attempt to hide social relations, to push cultural issues under the rug, to intentionally mislead the public about the nature of agents, etc. Rather, I believe that because AI is part of the scientific and engineering traditions, most alternativists simply do not have the training to include these aspects in their work. In Chapter 3, I noted that science values simplification through separation, and one of the key ways in which this is done is by separating the object of study from the complex and rich life background in which it exists. This strategy lets researchers focus on and hopefully solve the technical problems involved without getting bogged down in all kinds of interconnected and complex issues which may not have direct bearing on the task at hand.

The Return of the Repressed

The problem, though, is that even from a straightforward technical point of view, excluding the sociocultural context is sometimes unhelpful. At its most basic, ignoring this context does not make it go away. What ends up happening is that, by insisting that cultural influences are not at work, those influences often come back through the back door in ways that are harder to understand and utilize.

As an example, consider the use of programming through the use of symbols. Symbolic programming involves the use of tokens, often with names like “reason,” “belief,” or “feeling” which are loaded with cultural meaning to the agent designer. Critics point out that the meaningfulness of these terms to humans can obscure the vacuousness of their actual use in the program. So a programmer who writes a piece of code that manipulates tokens called ‘thoughts’ may unintentionally lead him- or herself into believing that this program must be thinking.

Alternative AI, generally speaking, involves a rejection of these sorts of symbols as tokens in programs. This rejection is often based on a recognition that symbolic programming of the kind classical AI engages in is grounded in culture, and that symbols carry a load of cultural baggage that affects the way programs are understood. Some of them believe that by abandoning symbolic programming they, unlike classicists, have also abandoned the problem of cultural presuppositions creeping into their work. And, in fact, it is true that many alternative AI programs do use such symbols sparingly, if at all, in their internal representations.

Nevertheless, it would be fair to say that the architecture of such agents involves symbols *to the extent that the engineer of the agent must think of the world and agent in a symbolic way in order to build the creature*. For example, the creature may have more or less continuous sensors of the world, but each of those sensors may be interpreted in a way that yields, once again, symbols — even when those symbols are not represented explicitly as a written token in an agent’s program. For example, a visual image may be processed to output one of two control signals, one of which triggers a walking style appropriate when on carpets, and one of which triggers a walking style appropriate when

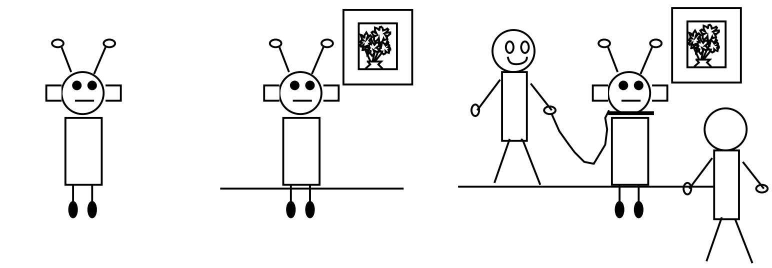

FIGURE 4.1: The increased context from classical through alternative to socially situated AI.

not on carpets. While a variable named 'on-carpet' may not appear in the agent's code, it would be fair to predicate an 'on-carpet' symbol *in the designer's thinking* as s/he constructed the agent - a symbol which is as informed by the designer's cultural background as the identifiable 'on-carpet' symbol in a classical AI program.

The behaviors into which the agent is split up are similarly fundamentally symbolic ("play fetch," "sleep," "beg," etc.) and are influenced by cultural notions of what behaviors can plausibly be. While alternative AI has gotten away from symbolic representations within the agent when seen in isolation, it has not gotten away from symbolic representations when the agent is seen in its full context. Once you look at the entire environment of the agent, including its creator, it is clear that despite the rhetoric that surrounds alternative AI, these symbols — and their accompanying sociocultural baggage — still play a large role.

Leaving out the social context, then, is both epistemologically inadequate and obfuscating. By not looking at the subjective aspects of agent design, the very nature of alternative AI programming, as well as the origin of various technical problems, becomes obscured. This is particularly problematic because not being able to see what causes technical problems may make them hard, if not impossible, to solve. We will see in the next chapter that this is exactly what happens with schizophrenia — and that by taking the opposite tack a path to solution becomes much more straightforward.

Socially Situated AI

What should AI do instead? Alternativists believe that situating agents in their physical context often provides insight into otherwise obscure technical problems. I propose that we build on this line of thinking by taking seriously the idea that the social and cultural environment of the agent can also be, not just a distracting factor in the design and analysis of agents, but a valuable resource for it (Figure 4.1. I coined the term 'socially situated AI' for this method of agent research.

Here, I will first describe at a philosophical level the postulates of socially situated AI. This lays out the broad framework within which technical work can proceed. I'll then discuss at a more concrete level what it means to design and build agents with respect to their sociocultural context. This concrete description will form the basis for redefinition of schizophrenia in the next chapter.

Postulates of Socially Situated AI

Like other methodological frameworks, including classical and alternative AI, socially situated AI involves, not just a kind of technology, but a way of understanding how to define problems and likely avenues of success. I represent this changed way of thinking here through an enumeration of postulates of socially situated AI. These are propositions that form the framework for how research is done and evaluated. Specifically, socially situated AI distinguishes itself from other forms of AI through explicit commitment to the following principles:

1. *An agent can only be evaluated with respect to its environment, which includes not only the objects with which it interacts, but also the creators and observers of the agent.* Autonomous agents are not 'intelligent' in and of themselves, but rather with reference to a particular system of constitution and evaluation, which includes the explicit and implicit goals of the project creating it, the group dynamics of that project, and the sources of funding which both facilitate and circumscribe the directions in which the project can be taken. An agent's construction is not limited to the lines of code that form its program but involves a whole social network, which must be analyzed in order to get a complete picture of what that agent is, without which agents cannot be meaningfully judged.

2. *An agent's design should focus, not on the agent itself, but on the dynamics of that agent with respect to its physical and social environments.* In classical AI, an agent is designed alone; in alternative AI, it is designed for a physical environment; in socially situated AI, an agent is designed for a physical, cultural, and social environment, which includes the designer of its architecture, the creator of the agent, and the audience that interacts with and judges the agent, including both the people who engage it and the intellectual peers who judge its epistemological status. The goals of all these people must be explicitly taken into account in deciding what kind of agent to build and how to build it.

3. *An agent is a representation.* Artificial agents are a mirror of their creators' understanding of what it means to be at once mechanical and human, intelligent, alive, what cultural theorists call a subject. Rather than being a pristine testing-ground for theories of mind, agents come overcoded with cultural values, a rich crossroads where culture and technology intersect and reveal their co-articulation. This means in a fundamental sense that, in our agents, we are not *creating* life but *representing* it, in ways that make sense to us, given our specific cultural backgrounds.

Socially Situated AI as Technical Methodology

These philosophical principles do not necessarily give technical researchers much to go on in their day-to-day work. Concretely speaking, socially

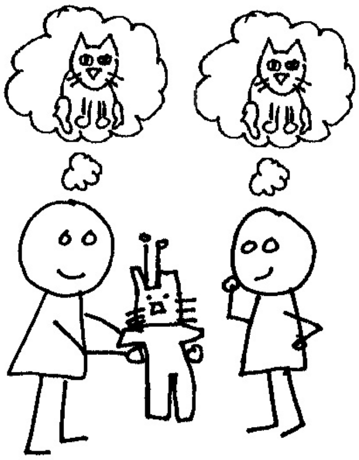

FIGURE 4.2: Agents as communication.

situated AI can be understood in the following way. Rather than seeing an agent as a being in a social vacuum, we can see it as represented in Figure 4.2: as a kind of *communication* between a human *designer* who is using it to embody a conception of an agent and a human *audience* who is trying to understand it.

After all, for many applications it is not enough for an agent to function correctly in a technical sense. Many times, the agent should also be *understandable*. For example, when an agent researcher designs an artificial cat, s/he will have some ideas about the kinds of behaviors the cat should have and the kind of motivations behind the cat's selection of various behaviors — ideas which, optimally and sometimes crucially, the viewers of the agent should also pick up on. In this sense the agent as program is a kind of vehicle for a conception of a particular agent, which is communicated from the agent-builder through the technical artifact to the observers of or interactors with the agent.

This way of understanding socially situated AI can be thought of as a change in metaphor. Many current approaches to AI are based on the metaphor of *agent-as-autonomous*: the fundamental property of such an agent is its basic independence from its creator or users. Lenny Foner, for example, defines autonomy as one of the most basic aspects of being an agent.

> Any agent should have a measure of autonomy from its user. Otherwise, it's just a glorified front-end, irrevocably fixed, lock-step, to the actions of its user. A more autonomous agent can pursue agenda independently of its user. This requires aspects of periodic action, spontaneous execution, and initiative, in that the agent must be able to take preemptive or independent actions that will eventually benefit the user. [Foner, 1993]

This autonomy implies that the agent's fundamental being is as a thing-for-itself, rather than what it actually is: a human construction, usually a tool. AI researchers are far from believing that agents magically spring from nowhere, and autonomy can certainly be a useful notion. Nevertheless, the focus on autonomy — separation from designer and user — as a *defining* factor for agents can unwittingly hide the degree to which both designers and users are involved in the agent's construction and use.

As an alternative to this metaphor, socially situated AI suggests the metaphor of *agent-as-communication*. Socially situated AI sees agents not as beings in a vacuum, but as representations which are to be communicated from an agent-builder to an audience. This point of view is deeply informed by recent work in believable agents such as [Neal Reilly, 1996] [Loyall, 1997a] [Wavish and Graham, 1996] [Blumberg and Galyean, 1995], which focus more and more on the audience's perception of agents, rather than on an agent's correctness per se. This conception of agents is also very like contemporary conventional conceptions of artwork, as vehicles through which ideas can be transmitted from a designer to his or her audience.

But the concept of agent-as-communication is not limited to believability or other 'artsy' applications. This is because proper perception of agents matters not only when we want to communicate a particular personality through our agents. It matters in *any* situation where the design of the agent — including its purpose, methods, functions, or limitations — should be understood by the people with which the agent interacts.

Thinking of agents as communication has several advantages. By making the commitment that 'agentiness' is meant to be communicated, we can explicitly communicate to the audience what the agent is about, rather than assuming (often incorrectly) that this will happen as a side-effect of the agent "doing the right thing." And by building agents with an eye to their reception, builders can tailor their agents to maximize their effectiveness for their target audience. In this sense, agents built for social contexts can be not only more engaging but more *correct* than purely rational, problem-solving agents, in the following sense: they may actually get across the message for which they have been designed.

This change in metaphor from autonomy to communication will have crucial implications, both in redefining the problem of schizophrenia in the next chapter, and for agent architecture down to its very details, as we will see later. It will turn out that behavior-based technology is so heavily invested in the metaphor of agent-as-autonomous that switching to agent-as-communication will have ramifications throughout the agent architecture. In the next chapter, we will see that taking seriously the quality of agent communication means redefining even the basic building blocks of behaviors as *signifiers*. In Chapter 7, we will learn that communication of agent motivation necessitates the use of *transitions* to explain to the user the agent's normally implicit decision-making process. But before we get to these changes, we will go back to the technical problem of schizophrenia with which we started, and look at how socially situated AI redefines the relationship between schizophrenia and atomization, showing us a way out of the conundrums of Chapter 2.

Chapter 5

Architectural Mechanisms I: Transitions as De-Atomization

Socially situated AI as a theory is well and good, but the proof of the pudding is in whether it actually helps us do anything differently. This chapter is devoted to exploring the technical consequences of the theoretical framework we have been developing in the last 2 chapters. Conceptually, we will start by rethinking the technical problem of schizophrenia as defined in Chapter 2 from the vantage point gained by the forays we have made into humanism. This new conception immediately suggests that the problem of schizophrenia should be rephrased. Instead of looking at schizophrenia as a property of agent code, we will look at schizophrenia as a problem of agent communication.

This way of rephrasing of the technical problem is amenable to more-or-less straightforward technical solution. I will use conventional AI techniques to solve this problem, leading to the following architectural innovations:

1. Behaviors are re-understood as *signifiers*, which explicitly act to communicate the agent's activity to users through the use of low-level *signs*. A *sign-management system* allows the agent to keep track of which signs and signifiers have been communicated to the user, so that the agent can make behavioral decisions based not only on what *it* thinks it is doing, but also on the likely user impression of its activities so far.

2. Sudden breaks between these signifying behaviors are smoothed over using *transitions*. Instead of leaping from behavior to behavior in the manner of the schizophrenic agents of Chapter 2, the agent gradually morphs between them.

3. These transitions are implemented using *meta-level controls*, which allow behaviors to share information and coordinate their effects. By making the coordination of behaviors explicit — rather than an implicit side-effect of the underlying architecture — meta-level

controls allow the relationships between behaviors to be expressly communicated to the user.

If you are not technically trained this chapter can be viewed as a case study in AI methodology. Since the rhetorical style of AI argumentation is not always transparent to those not trained in AI, margin boxes will provide some context by explaining the role of each piece in developing the larger argument. AI researchers may also find this outsider perspective on AI argumentation enlightening!

It will turn out that this basically purely technical approach works to smooth observable behavior together, thereby making agents seem less schizophrenic. Unfortunately, that in itself will not necessarily help make agents that are effective in appearing truly intentional. To put it simply, the techniques developed here may keep the agents from looking transparently bad (which is of course nice), but they don't necessarily make them look particularly good. For that, we will need to think more deeply about the assumptions and requirements of the technical approach. We will do this through another foray into animation (Intermezzo II) and psychology and the cultural studies of science (Chapter 6). These will allow us to build on the technical foundations of this chapter to create the full agent architecture in Chapter 7.

Socially Situated AI vs. Good Old-Fashioned Alternative AI

The technical developments in this chapter depend in a deep sense on understanding how socially situated AI fundamentally changes the ground on which alternative AI operates. As discussed in the previous chapter, socially situated AI suggests that the agent and its behavior should be thought about, not in terms of the agent itself, but in terms of communication between the designer of the agent and its audience. Rather than *intelligent* agents, then, the focus is on creating *intelligible* agents, ones that successfully communicate to the audience the idea for the agent that the designer had in mind.

This switch from intelligence to intelligibility may be recognizable to AI researchers as the mindset change behind believable agents that motivates such work as [Bates, 1994], [Loyall, 1997a], and [Neal Reilly, 1996]. Believable Agents — characters that are intended to communicate a particular artist-chosen personality — similarly focus on situated communication over an agent's abstract (and perhaps uncommunicated) reasoning abilities. Socially situated AI builds on a rich foundation laid by Believable Agent researchers, by seeing this communication perspective as not only useful for agents that are to inhabit works of art or entertainment, but for *all* agents — whether intended as living creatures or as helpful tools — whose activity should be comprehensible to humans with which it interacts. This may include agents like office robots, tele-autonomous systems, or automated flight systems, whose *function* is totally utilitarian, but whose actions should be understandable in order to function well with and to inspire confidence from the humans who come into contact with them.