Schizophrenia and Narrative in Artificial Agents

Contact
Phoebe Sengers
Media Arts Research Studies
Institüt für Medienkommunikation
GMD Forschungszentrum
Informationstechnik GmbH
Schloss Birlinghoven
D-53754 Sankt Augustin, Germany
phoebe.sengers@gmd.de

In recent years, computer graphics has turned to AI techniques in order to simplify the problem of modeling moving objects for rendering.[4,19,21] By modeling the minds of graphically represented creatures, their movements can be directed automatically through AI algorithms and need not be directly controlled by the designer. But what kind of baggage do these AI algorithms bring with them? Here I will argue that predominant AI approaches to modeling agents result in behavior that is fragmented, depersonalized, lifeless, and incomprehensible. Drawing inspiration from narrative psychology and anti-psychiatry, I will argue that agent behavior should be narratively understandable and present an agent architecture that structures behavior to be comprehensible as narrative.

The approach I take in this essay is a hybrid of critical theory and AI agent technology. It is one example of a critical technical practice [1]: a cultural critique of AI practice instantiated in a technical innovation. In the final section of this paper, I will describe the theoretical and practical foundations of the critical technical practice pursued here, which I term socially situated AI.

Introduction

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

Roboticist Rodney Brooks expresses this worry eloquently:

Perhaps it is the case that all the approaches to building intelligent systems are just completely off-base, and are doomed to fail.... [C]ertainly it is the case that all biological systems.... [b]ehave in a way which just simply seems "life-like" in a way that our robots never do.

Perhaps we have all missed some organizing principle of biological systems, or some general truth about them. Perhaps there is a way of looking at biological systems which will illuminate an inherent necessity in some aspect of the interactions of their parts that is completely missing from our artificial systems.... [P]erhaps at this point we simply do not get it, and... there is some fundamental change necessary in our thinking... [P]erhaps we are currently missing the juice of life.[6]

Here, I argue that the "juice" we are missing is narrative. The divide-and-conquer methodologies currently used to design artificial agents result in fragmented, depersonalized behavior, which mimics the fragmentation and depersonalization of schizophrenia

in institutional psychiatry. Anti-psychiatry and narrative psychology suggest that the fundamental problem for both schizophrenic patients and agents is that observers have difficulty understanding them narratively. This motivates a narrative agent architecture, the Expressivator, which structures agent behavior to support narrative, thereby creating agents that are intentionally comprehensible.

The Problem

Building complex, integrated artificial agents is one of the dreams of AI. Classically, complex agents are constructed by identifying functional components (natural language processing, vision, planning, etc.), designing and building each separately, then integrating them into an agent. More recently, some practitioners have argued that the various components of an agent strongly constrain one another, and that the complex functionalities that classical AI could come up with could not easily be coordinated into a whole system. They offer other construction methodologies instead. In particular, behavior-based AI proposes that the agent should be split up, not into disparate cognitive functionalities, but into behaviors, each of which integrates all of the agent's functions for a particular behavior in which the agent engages. Examples of such behaviors include foraging, sleeping, and hunting.

Even such systems, however, have not been entirely successful in building agents that integrate a wide range of behaviors. Rod Brooks, for example, has stated that one of the challenges of the field is to find a way to build an agent that can integrate many behaviors, and he defines many to be more than a dozen.[5] Programmers can create robust, subtle, effective, and expressive behaviors, but the agent's overall behavior tends to gradually fall apart as more and more behaviors are combined. For small numbers of behaviors, this disintegration can be managed by the programmer, but as more and more behaviors are combined, their interactions become so complex that they become at least time-consuming and at worst impossible to manage.

In both cases, divide-and-conquer methodologies lead to integration problems. With classical agents, who are split up by functionality, there are often problems with a functional under-integration. This underintegration manifests itself in various kinds of inconsistency between the different functions, such as not being able to use knowledge for one function that is available for another. For example, the agent may speak a word it cannot understand or visibly register aspects of the world that do not affect its subsequent behavior. In behavior-based agents, under

integration manifests itself on the behavioral level. These agents generally have a set of black-boxed behaviors. Following the action-selection paradigm, agents continuously redecide which behavior is most appropriate. As a consequence, they tend to jump around from behavior to behavior according to which one is currently the best.[28]

What this means is that the overall character of the agent's behavaior ends up being deficient; generally speaking, its behavior consists of short dalliances in individual, shallow, high-level behaviors with abrupt changes between behaviors. It is this overall defective nature of agent behavior, caused by under-integration of behavioral units, that I term schizophrenia and propose to address here.

Schizophrenia is a loaded term. I use it here to draw attention to important connections between current approaches to agent-building and the experience of being schizophrenic in institutional psychiatry. In the next two sections, I draw out those connections, then show how an alternative approach to psychiatric schizophrenia can motivate changes in AI practice. These changes form the basis for narrative agent architecture.

Schizophrenia

Schizophrenia's connection to AI is grounded in one of its more baffling symptoms: the sentimente d'automatisme, or subjective experience of being a machine.[12] 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."[23] R. D. Laing describes how some schizophrenic patients experience or fear experiencing themselves as things, as its, instead of as people.[13] 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.[13]

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 section: "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.[13] 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.[13]

In this sense, there is a direct link between schizophrenia and behavior-based methodology – and symptomatology.

Depersonalization

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, as a number of critics have argued, psychiatric institutions themselves reinforce this feeling of mechanicity and lack of autonomous self. For example, Erving Goffman, in his ground-breaking anthropological study "Asylums",[11] argues that a major feature of psychiatric institutions is the "programming" of each inmate "into an object that can be fed into the administrative machinery of the establishment, to be worked on smoothly by routine operations."[11]

One of the signs of this mechanization is the reduction of patient to symptomatology. Patients are constantly monitored, their behavior continuously examined for and interpreted as signs of illness. 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. Rather than being taken seriously as such, a patient's 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.[22]

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.[2]

The patient is formalized, reduced to a set of somewhat arbitrarily connected symptoms. 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 fragmentation into symptoms, Laing argues, actually reinforces, rather than treats, schizophrenia. When mechanistic explanations reduce the patient to a bundle of pathological processes, 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.... [W]e are [then] 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.[13]

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

This problem of conceptual splitting parallels closely the problem of AI, suggesting that mechanistic explanations of the sort necessary to build agents are also responsible for their de-intentionalized appearance. The symptomatology of institutional psychiatry is reflected in behavioral black-boxing in behavior-based AI. In the next section, we will explore alternatives to this fragmentation in psychiatry, searching for clues for dealing with the problem of schizophrenia in AI.

ANTI-PSYCHIATRY

In the 1960s and 1970s, Laing and other sympathetic colleagues, termed anti-psychiatrists for their opposition to mainstream psychiatry, suggested that the schizophrenizing aspects of institutional psychiatry can be avoided by changing our viewpoint on patients: instead of thinking of schizophrenics as self-contained clusters of symptoms, we should try to understand them phenomenologically, as complex humans whose behavior is meaningful. The schizophrenizing 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.[14]

The phenomenological approach, 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.[14]

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.[14]

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 she what she thinks.

This focus on hermeneutic interpretation rather than data extraction as a way of understanding intentional behavior can be applied to agent design. From this perspective, when we focus largely on the decomposition of agents' behavior into individually designed units, we will necessarily end up with fragmented and depersonalized agents. On the other hand, if we take an interpretive, wholistic perspective to agents, we may be able to build agents without undermining their intentionality. In solving the problem of schizophrenic agents, this is a lead, but only that. In order to make concrete changes in agent technology, we need to have a more exact understanding of what this change in"intentional stance"[9] means. We will use narrative psychology to specify the change in understanding suggested by anti-psychiatry; this, it turns out, will give us a toehold in agent design.

NARRATIVE PSYCHOLOGY

 arrative psychology, an area of study developed by Jerome Bruner,[7,8] focuses on how people interpret specifically intentional behavior. Narrative psychology shows that, whereas people tend to understand inanimate objects in terms of cause-effect rules and by using logical reasoning, intentional behavior is made comprehensible by structuring it into narrative or "stories." Narrative psychology suggests that this process of creating narrative is the fundamental difference between the way people understand intentional beings and mechanical artifacts.

That is to say, if I want to understand and build an inanimate object, I may decompose it, try to understand what different pieces are for, replicate how they work, and figure out the rules underlying its behavior. On the other hand, if I want to understand a person's behavior, I am interested in such things as what motivates him or her, the reasons he or she engages in particular activity, and how his or her behavior reflects on his or her whole personality.

This contrast between narrative explanations that explore the meaning of living activity and atomistic explanations that allow for the understanding and construction of mechanical artifacts provides a theoretical basis for the criticisms of anti-psychiatry. Anti-psychiatrists, after all, complain that the difficulty with institutional psychiatry is that it reduces the patient to a pile of data, thereby making a machine of a living person. The anti-psychiatric solution of interpretation uses narrative understanding to "repersonalize" patients: structuring and relating the "data" of a patient's life into the semi-coherent story of a meaningful, though painful, existence; focusing on the patient not as an instance of a disease but as a particular individual and how that person feels about his or her life experience; and relating the doctor's narrative to its background conditions and the life context in which it is created and under-

stood. It is only through this process of narrative interpretation that, according to anti-psychiatry, the psychiatrist can fully respect and understand the patient's subjective experience as a human being.

In AI, this distinction between mechanism and intentional being becomes problematic. AI agents should ideally be understandable both as well-specified physical objects and as sentient creatures. In order to understand intentional behavior, users attempt to construct narrative explanations of what the presumed intentional being is doing, but this approach conflicts with the mechanistic explanations that designers themselves need to use in order to identify, structure, and replicate behavior. The resulting abrupt behavioral breaks create the (often correct) impression that there is no relationship between the agent's behaviors; rather than focusing on understanding the agent as a whole, the user is left to wonder how individually recognizable behaviors are related to each other and the agent's personality. Behaviors are designed in isolation and interleaved according to opportunity, but users, like it or not, attempt to interpret behaviors in sequence and in relationship to each other. The result of this mismatch between agent design and agent interpretation is confusion and frustration on the part of the user and the destruction of apparent agent intentionality.

At this point, there seems to be a basic and unsolvable mismatch between fragmentation and intentionality. But narrative psychology suggests that the fundamental problem with current agent-building techniques is not simply recognizable fragmentation in and of itself, but rather that fragmented agents do not provide proper support for narrative interpretation. From this follows the major insight of this paper: If humans understand intentional behavior by organizing it into narrative, then our agents will be more"`intentionally comprehensible" if they provide narrative cues. That is to say, rather than simply presenting intelligent actions, agents should give visible cues that support users in their ongoing mission to generate narrative explanation of an agent's activity. We can do this by organizing our agents so that their behavior provides the visible markers of narrative.

NARRATIVE AGENT ARCHITECTURE

What does it mean for agents to support narrative comprehension? The properties of narrative are complex.[8] Elsewhere, I have discussed in detail how they can apply to AI.[25,26] For the sake of brevity, I will here limit discussion to the following properties:

Context-Sensitivity and Negotiability

In behavior-based systems, the "meaning" of a behavior is thought of as always the same: the name the designer gives the internally-defined behavior. But in narrative comprehension, meaning is not a matter of identifying already-given symbols, but comes out of a complex process of negotiation between the interpreter and the events being interpreted. The meaning of the same event can change radically based on the context in which it occurs, as well as on the background, assumptions, knowledge, and perspective of the interpreter. In order to design narratively expressive agents, designers must respect (rather than attempt to override) the context- and audience-dependency of narrative comprehension.

Intentional State Entailment

In most behavior-based systems, the reason a behavior is run is implicit in its action-selection mechanism. The behavior is then necessarily communicated to the user on a "just-the-facts-ma'am" basis: it is usually easy to see what an agent is doing, but hard to tell why. But in narrative, the reasons or motivations behind actions are just as important as (if not more important than) what is done. People do not want to know just the events that occur in the narrative. They also want to know the motivations, thoughts, and feelings behind them. Supporting narrative comprehension means communicating clearly not just what the agent does, but its reason for doing it.

Diachronicity

Behavior-based agents jump from behavior to behavior according to what is currently optimal. Each of these behaviors is designed independently, with minimal interaction. But a fundamental property of narrative is its diachronicity; it relates events over time. In a narrative, events do not happen randomly and independently; they are connected to and affect one another. Narrative support in a behavior-based agent requires normally independent behaviors to be able to influence each other, to present a coherent picture of narrative development to the user over time.

These properties are the motivation for the Expressivator, an agent architecture that focuses on the narrative expression of agent behavior. The Expressivator is an extension of Bryan Loyall's Hap,[15,16] a behavior-based language designed for believable agents. The Expressivator has been tested in the Industrial Graveyard, a virtual environment in which the Patient, a discarded lamp character implemented with the Expressivator, attempts to eke out a miserable existence while being bullied about by the Overseer, an agent implemented in Hap. Generally, the Expressivator supports narrative comprehension using the following heuristic:

Behaviors should be as simple as possible. The agent's life comes from thinking out the connections between behaviors and displaying them to the user.

Simpler behaviors are essential because complex processing is lost on the user. Most of the time, the user has a hard time picking up on the subtle differences in behavior that bring such pleasure to the heart of the computer programmer. But the properties of narrative

interpretation mean that simpler behaviors are also enough. Because the user is very good at interpretation, minimal behavioral cues suffice.

More specifically, the Expressivator provides systematic support for narrative comprehensibility through the following mechanisms:

Context-Sensitivity and Negotiability

Rather than building an agent from conventional context- and communication-independent actions and behaviors, a designer builds agents from context-dependent signs and signifiers that are to be communicated to the user.

Intentional State Entailment

Transitions are added between signifiers to explain why the agent's observed behavior is changing.

Diachronicity

Signifiers can use meta-level controls to influence one another, presenting a coherent behavioral picture over time.

SIGNS, SIGNIFIERS, AND SIGN MANAGEMENT

Typical behavior-based agents are designed for correctness, not for user comprehensibility. The first step the Expressivator takes in creating narratively understandable agents is to open the architecture up for communication. Agent design is based, not on the functions the agent must fulfill, but on its intended, context-dependent interpretation by the user. In the Expressivator, signs and signifiers support the construction of clearly communicated behavior; sign management allows the agent itself to keep track of what has been communicated, so it can tailor subsequent behavioral communication to the user's current interpretation.

Signs and Signifiers

Current behavior-based approaches are based on an internal, problem-solving approach, and generally divide an agent into activities in which the agent likes to or needs to engage. Typical behavior-based systems divide an agent into three parts:

1. Physical actions in which the agent engages
2. Low-level behaviors, which are the agent's simple activities
3. High-level behaviors, which combine low-level behaviors into high-level activities using more complex reasoning.

Because these activities are implemented according to what makes sense from the agent's internal point of view, there is no necessary correlation between the agent's behaviors and the behaviors we would like the user to see in our agent. But if the agent is to be narratively comprehensible, it may make more sense to design the agent according to the desired user interpretation. We may want to make the internal behaviors exactly those behaviors we want to communicate to the user. Then, communicating what the agent does reduces to the problem of making sure that each of these behaviors is properly communicated. For this reason, the Expressivator structures an agent not

68

into physical actions and problem-solving behaviors, but into signs and signifiers, or units of action that are likely to be meaningful to the user. This structure involves three levels, roughly corresponding to those of generic behavior-based AI:

1. Signs, which are small sets of physical actions that are likely to be interpreted in a particular way by the user
2. Low-level signifiers, which combine signs, physical actions, and mental actions to communicate particular immediate physical activities to the user.
3. High-level signifiers, which combine low-level signifiers to communicate the agent's high-level activities.

There are several differences between these structural units and the default behavior-based ones. Unlike physical actions and behaviors, signs and signifiers focus on what the user is likely to interpret, rather than what the agent is "actually" (internally) doing. In addition, signs and signifiers are context-dependent; the same physical movements may lead to different signs or signifiers, depending on the context in which the actions are interpreted. Most importantly, signs and signifiers carry an explicit commitment to communication; they require the agent designer to think about how the agent should be interpreted and to provide visual cues to support that interpretation.

Signs and signifiers are not simply design constructs; they also have technical manifestations. Formally, a sign is a token the system produces after having engaged in physical behavior that is likely to be interpreted in a particular way. This token consists of an arbitrary label and an optional set of arguments. The label, such as "noticed possible insult," is meaningful to the designer and represents how the designer expects that physical behavior to be interpreted. The arguments (such as "would-be insulter is Wilma") give more information about the sign. This token is stored by the sign-management system described below, so that the agent can use it to influence its subsequent behavioral decisions. A low-level signifier is a behavior that is annotated with the special form "(with low_level_signifying...);" a high-level signifier is similarly annotated "(with high_level_signifying....)." Signifiers can also generate tokens for the sign-management system, as described below.

Sign Management

Once a designer has structured an agent according to what it needs to communicate, agents can reason about what has been communicated in order to fine-tune presentation of subsequent signs and signifiers. That is, by noting which signifiers have been communicated, agents can reason about the user's likely current interpretation of their actions and use this as a basis for deciding how to communicate subsequent activity. The most obvious way for the agent to keep track of what the user thinks is for it simply to notice which signs and signifiers are currently running. After all, signifiers represent what is being communicated to the user. But in

practice, it turns out that this is not correct because the user's interpretation of signs and signifiers lags behind the agent's engagement in them. For example, if the agent is currently running a "head-banging" signifier, the user will need to see the agent smack its head a few times before realizing that the agent is doing it.

The sign-management system deals with this problem by having the agent post signs and signifiers when it believes the user must have seen them. A behavior can post a sign each time it has engaged in some physical actions that express that sign, using the "post_sign" language mechanism. Similarly, once signs have been posted that express a low-level signifier, behaviors use "post_low_level" to post that that low-level signifier has been successfully expressed. Once the right low-level signifiers have been posted to express a high-level signifier, "post_high_level" is used to post that high-level signifier.

Each of these commands causes a token to be stored in the agent's memory listing the current sign, low-level signifier, or high-level signifier, respectively, along with a time stamp. Once signs and signifiers have been posted, other behaviors can check to see what has been posted recently before they decide what to do. The result is that the signs and signifiers the agent has expressed can be used just like environmental stimuli and internal drives to affect subsequent behavioral presentation, tuning the agent's behavior to the user's interpretation.

Transitions

The second requirement of narrative comprehensibility is that the user should be able to tell why the agent is doing what it is doing. In behavior-based terms, every time an agent selects a particular behavior, it should express to the user the reason it is changing from the old behavior to the new one. This is difficult to do in most behavior-based systems because behaviors are designed and run independently; when a behavior is chosen, it has no idea who it succeeds, let alone why.

In the Expressivator, behavioral transitions are used to express the agent's reasoning. Transitions are special behaviors which act to "glue" two signifying behaviors together. When a transition notices that it is time to switch between two signifiers, it takes over from the old signifier. Instead of switching abruptly to the new signifier, it takes a moment to express to the user the reason for the behavioral change.

Transitions are implemented in two parts, each of which is a full-fledged behavior: transition triggers, which determine when it is appropriate to switch to another behavior for a particular reason, and transition demons, which implement the transition sequence that expresses that reason to the user. Transition triggers run in the background, generally checking which behaviors are running (e.g., exploring the world), and combining this information with sensory input about current conditions (e.g., the Overseer is approaching). When its conditions are fulfilled, the transition trigger adds a special token to memory, noting the behavior that should terminate, the behavior that should replace it, and a label that represents the reason for the replacement (e.g., "afraid_of_overseer"). Transition demons monitor memory, waiting for a transition for a particular reason to be triggered. They then choose an appropriate behavioral expression for the reason for change, according to the current likely user interpretation and conditions in the virtual environment. Expressing the reasoning behind behavioral change often requires changes to subsequent behaviors; for example, if the Patient starts doing some odious task because it is forced to by the Overseer, it should include some annoyed glances at the Overseer as part of the task-fulfilling behavior. Transitions are able to express these kinds of interbehavioral influences using the meta-level controls described below.

Meta-Level Controls

The third requirement of narrative comprehensibility is that behaviors should be structured into a coherent sequence. Instead of jumping around between apparently independent actions, the agent's activities should express some common threads. But these relationships between behaviors are difficult to express in most behavior-based systems because they treat individual behaviors as distinct entities that do not have access to each other. Conflicts and influences between behaviors are not handled by behaviors themselves but by underlying mechanisms within the architecture. Because the mechanisms that handle relationships between behaviors are part of the implicit architecture of the agent, they are not directly expressible to the user.

The Expressivator deals with this problem by giving behaviors meta-level controls, special powers to sense and influence each other. Because meta-level controls are explicitly intended for communication and coordination between behaviors, they are in some sense a violation of the behavior-based principle of minimal behavioral interaction. Nevertheless, meta-level controls are so useful for coordinating behavior that several have already found a home in behavior-based architectures. An example is Hamsterdam's meta-level commands, which allow non-active behaviors to suggest actions for the currently dominant behavior to do on the side.[3] In the Expressivator, behaviors can:

1. Query which other behaviors have recently happened or are currently active
2. Delete other behaviors.
3. Add new behaviors, not as subbehaviors, but at the top level of the agent.
4. Add new sub-behaviors to other behaviors.

5. Change the internal variables that affect the way in which other behaviors are processed.
6. Turn off a behavior's ability to send motor commands.
7. Move running subbehaviors from one behavior to another.

The most important function for these meta-level controls in the Expressivator is to allow for implementation of transitions. Transitions, at a minimum, need to be able to find out when an old behavior needs to be terminated, delete the old behavior, engage in some action, and then start a new behavior. This means that transition behaviors need to have all the abilities of a regular behavior, and a few more:

- They need to be able to know what other behaviors are running.
- They need to be able to delete an old behavior.
- They need to be able to begin a new behavior.

Ideally, they should also be able to alter the new behavior's processing to reflect how it relates to what the agent was doing before. In the Expressivator, transitions can do all these things with meta-level controls. More generally, meta-level controls make the relationships between behaviors explicit, as much a part of the agent design as behaviors themselves. They allow behaviors to affect one another directly when necessary, rather than making interbehavioral effects subtle side-effects of the agent design. Meta-level controls give agent builders more power to expose the inner workings of agents by letting them access and then express aspects of behavior processing that other systems leave implicit.

Putting It All Together

Narrative psychology suggests that narrative comprehension is context-sensitive, focuses on agent motivation, and seeks connections between events over time. The Expressivator supports comprehensibility by expressing the agent's actions with signs and signifiers, the reasons for agent activity with transitions, and the coherent threads through activities with meta-level controls.

These architectural mechanisms are described separately but used together in the agent design process, changing the way in which agents are designed. In a typical behavior-based system, an agent is defined in three major steps:

1. Deciding on the high-level behaviors in which the agent will engage
2. Implementing each high-level behavior, generally in terms of a number of low-level behaviors and some miscellaneous behavior to knit them together.
3. Using environmental triggers, conflicts, and other design strategies to know when each behavior is appropriate for the creature to engage in.

With the Expressivator, the choice and expression of these structural "units" for the agent is not enough; in order to support the user's comprehension, the designer must also give careful consideration to expressing the reasons for and connections between those units. These connections are designed and implemented with transitions, which alter the signifiers they connect into a narrative sequence. In practice, transitions are the keystone of the architecture, combining signifiers in meaningful ways through the use of meta-level controls.

Results

The best way to see how the Expressivator changes the quality of agent behavior is to look at how its transitions work in detail. Here, I will go over one point where the agent switches behaviors and explain how transitions make this switch more narratively comprehensible. One example does not proof make, but it does take up a lot of space; the sceptical reader can find more my earlier work.[25]

As our excerpt begins, the Patient notices the schedule of daily activities that is posted on the fence and goes over to read the schedule. The Overseer, noticing that the Patient is at the schedule and that the user is watching the Patient, goes over to the schedule, changes the time to 10:00, and forces the Patient to engage in the activity for that hour: exercising.

The goal of this part of the plot is to communicate to the user the daily regime into which the Patient is strapped. The Patient does not have autonomy over its actions; it can be forced by the Overseer to engage in activities completely independently of its desires. The specific behavioral change from reading the schedule to exercising, then, should show the user that the agent changes its activity because:

- It notices the Overseer.
- The Overseer enforces the scheduled activities
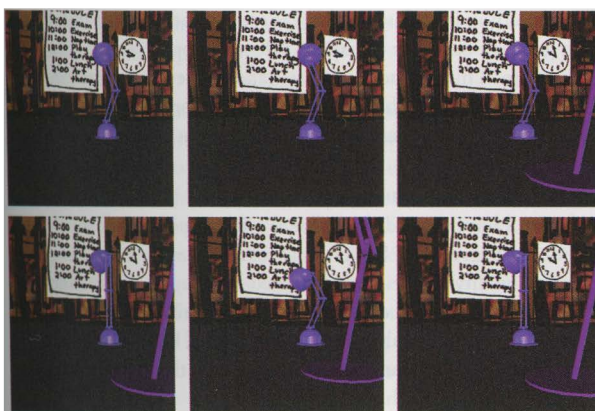- The activity that is currently scheduled is exercising.

Without transitions, the Patient's response to the Overseer is basically stimulus-response (Figure 2). The Patient starts out reading the schedule. As soon as the Patient senses the Overseer, it immediately starts exercising. This reaction is both correct and instantaneous; the Patient is doing an excellent job of problem-solving and rapidly selecting optimal behavior. But this behavioral sequence is also perplexing; the chain of logic that connects the Overseer's presence and the various environmental props to the Patient's actions is not displayed to the user. It is jumped over in the instantaneous change from one behavior to another.
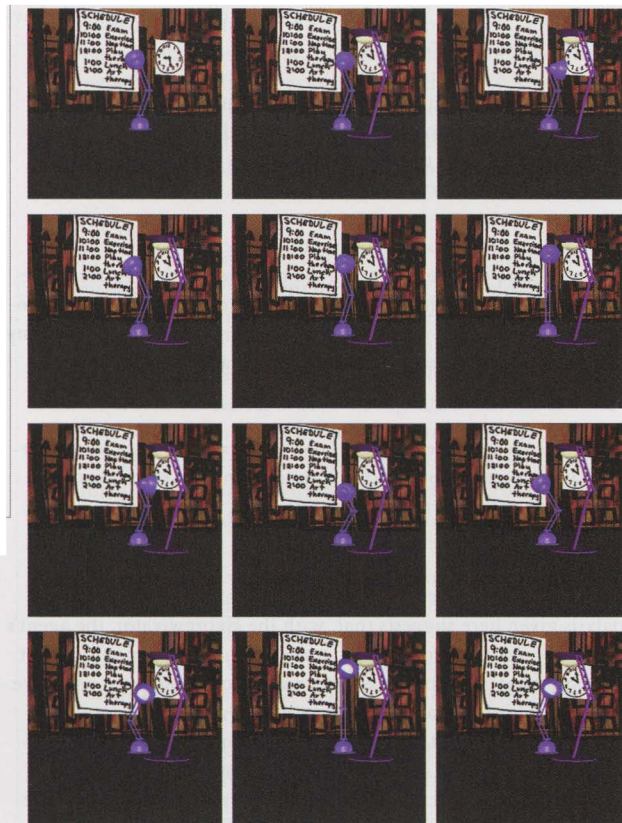

Figure 3. Response with transitions

Figure 2. Response without transitions

71

With transitions, attempts are made to make the reasons behind the behavioral change clearer (Figure 3). Again, the behavior starts with the Patient reading the schedule. This time, when the Overseer approaches, the Patient just glances at the Overseer and returns to reading. Since the Patient normally has a strongly fearful reaction to the Overseer (and by this time the Overseer's enthusiasm for punishing the Patient has already generally aroused sympathy in the user's mind), the user has a good chance of understanding that this simple glance without further reaction means that the Patient has not really processed that the Overseer is standing behind it.

Suddenly, the Patient becomes startled and quickly looks back at the Overseer again. Now, the user can get the impression that the Patient has registered the Overseer's presence. Whatever happens next must be a reaction to that presence. Next, the Patient checks the time and the schedule of activities to determine that it is time to exercise. Then the Patient whirls to face the Overseer and frantically and energetically begins exercising, tapering off in enthusiasm as the Overseer departs. This transition narrativizes the agent's behavior in the following ways:

- The agent design is predicated on the user's context-dependent interpretation (for example, that the user will interpret the agent's short glance at the Overseer differently now than earlier in the story.

- The transition communicates that the change in behavior is connected to several factors: the presence of the Overseer, the clock, and the schedule. This is in contrast with the transitionless sequence, in which there is no clear connection among any of the environmental factors and the Patient's behavioral change.

- The subsequent exercising behavior is altered to fit into a narrative sequence by making it more frantic in response to the agent's panic during the transition.

EVALUATION
How good is the Expressivator? The kind of detailed transition analysis given here suggests that, with the Expressivator, the agent's behavior is designed for context, provides more information about the reasons for agent behavior, and makes for a smoother narrative sequence. This is certainly a basis for improved narrative understanding, but does not necessarily imply actual improvement. In particular, the quality of the animation is not up to snuff, which means users sometimes have trouble interpreting the simple movements of the agent. All the innovations the Expressivator introduces are worthless if individual signs are not clearly animated; everything rests on the substantial animation problem of getting a sigh to look like a sigh and not like a cough or a snort. This problem is exacerbated when, as in Hap, there is a mind-body split, with the mind generating actions that are implemented autonomously by the body. The resulting divide between command

and execution makes accurate timing and therefore effective control of animation impossible. This problem of generating expressive animation, while not a straightforward "AI problem," must be addressed by any architecture that is going to implement graphically presented, comprehensible agents.

The Industrial Graveyard is an entertainment application, but the constructs of the Expressivator are not limited to believable agents. The concept of a narrative structure for behavior can be just as important for tele-autonomous robots, semi-autonomous avatars, or pedagogical agents. However, the Expressivator's focus on visible behavior and concrete action probably does not adequately support systems like automatic theorem provers that engage in complex, abstract reasoning.

The greatest conceptual problem with the Expressivator is the potential explosion of the number of transitions needed between signifiers; but this turned out not to be a problem in practice. For the Patient's eight high-level signifiers, there were only 15 transitions, and for the Patient's 16 low-level signifiers, there were only 25 transitions. There are several reasons for this. First of all, transitions are only needed between high-level signifiers, and between low-level signifiers that share the same high-level signifier – not between low-level signifiers in different high-level signifiers. This would be implemented, instead, with a transition between the respective high-level signifiers. I also cut out many transitions by writing several generic transitions, which could go from any behavior to a particular behavior. Most importantly, I found in practice that many of the possible transitions did not make practical sense because of the semantics of the behaviors involved.

The greatest advantage of the Expressivator for the behavior programmer is that it makes it much easier to handle interbehavioral effects. Coordination of multiple high-level behaviors is one of the major stumbling blocks of behavior-based architectures.[5] Since interbehavioral factors are implicit in the architecture, they are hard to control, leading to multiple behaviors battling it out over the agent's body, and hours of tweaking to get each behavior to happen when and only when it is supposed to. This is much easier to handle when behaviors can simply kill other behaviors that are not appropriate and when the trigger conditions for each behavior can be explicitly set.

SOCIALLY SITUATED AI
So far, I have argued that there is a fundamental lack in autonomous agents' behavior, which reduces their apparent intentionality. By being constructed in a fragmented manner, agents suffer a kind of schizophrenia, a schizophrenia that can be addressed, in analogy to anti-psychiatry, by making agents narratively understandable. In order to do this, I have built an agent architecture that combines:

- Redefinition of behaviors as signifiers and their reorganization in terms of audience interpretation

- The use of transitions to explain agent motivation, structuring user-recognized behaviors into narrative sequences.

• The use of meta-level controls to strategically undermine fragmentation of the agent's behaviors. Preliminary results are encouraging, but further work, preferably involving development of support for graphical presentation, will be necessary in order to fully evaluate the implications of and possibilities for the architecture.

More generally, if black-box behaviorism involves thinking of human life mechanically, reducing it to a matter of cause-effect, while narrative allows for full elucidation of meaningful intentional existence, then it seems likely that narrative, and by extension the humanities, for whom narrative is a modus operandi, can address meaningful human life in a way that an atomizing science simply cannot. If humans comprehend intentional behavior by structuring it into narrative, then AI must respect and address that way of knowing in order to create artifacts that stimulate interpretation as meaningful, living beings. This suggests that the schizophrenia we see in autonomous agents is the symptomatology of an overzealous commitment to mechanistic explanation in AI, a commitment that is not necessarily unhelpful (since it forms the foundation for building mechanical artifacts), but needs to be balanced by an equal commitment to narrative as the wellspring of intentionality.

In this final section, I will show that the focus on narrative communication to generate artificial beings that appear lifelike is part of a broader shift in view that comes about when AI is looked at from a cultural perspective. The resulting perspective I term socially situated AI, which shares close affinity with culturally oriented approaches taken by other AI researchers, notably Michael Mateas,[17] Simon Penny,[18] and Warren Sack.[24]

Introduction

To recap, the analysis in the first sections of this paper suggests that AI and institutionalization share properties that lead to schizophrenia. Both AI 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. Because these data are not related to one another or the context from which they spring, the result is the fragmentation of experience that cultural theorists term schizophrenia.

The conclusion of this argument is that, in order to address schizophrenia, we can take the opposite approach. Rather than seeing 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 the early sections of this paper 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 section, 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 approaching AI is, in retrospect, the key to solving schizophrenia by suggesting the redefinition of the problem of schizophrenia as a difficulty of agent communication rather than of internal agent structure.

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 (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.

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, and 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"[27] – 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 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. 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 that 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 believe 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 on carpets, and one of which triggers a walking style appropriate for walking on uncarpeted surfaces. 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 the agent is constructed – a symbol that 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 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. ●nce 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. This is exactly what happens with schizophrenia, and 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 design and analysis of agents, but also a valuable resource for it (Figure 4). I coined the term "socially situated AI" for this method of agent research.
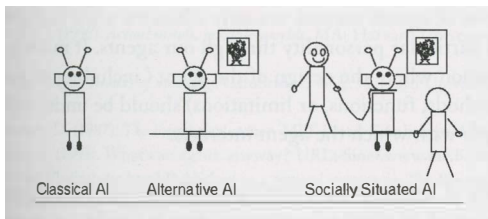
The increased context from classical through alternative to socially situated AI.

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.

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:

- 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. It 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.

- 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.

- 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, and 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 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 5, as a kind of communication between a human designer, who is using it to embody a conception of an agent. and a human audience that is trying to understand it.

Agents as communication.

75
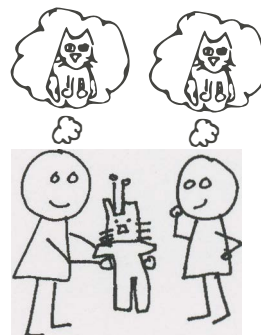
After all, for many applications it is not enough for an agent to function correctly in a technical sense. Many times, the agent should also be understandable. For example, when an agent researcher designs an artificial cat, the researcher will have some ideas about the kinds of behaviors the cat should 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 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 an 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.[10]

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 that are to be communicated from an agent builder to an audience. This point of view is deeply informed by recent work in believable agents,[20,15,29] that focuses 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. The notion of an agent as communication is clearly a more accurate description of how agents function culturally than the notion of an agent in an autonomous vacuum. It also brings advantages from a purely technical point of view. 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.

References

1. Philip E. Agre (1997). *Computation and human experience*. Cambridge, UK: Cambridge University Press, 1997.
2. Susan B. (1991). *The dinosaur man: Tales of madness and enchantment from the back ward*. New York: Edward Burlingame Books, 1991.
3. Blumberg, B. (1996). *Old tricks, new dogs: Ethology and interactive creatures*. PhD Thesis, MIT Media Lab, Cambridge, MA, 1996.
4. Blumberg, B. & Galyean, T. A. (1995). Multi-level direction of autonomous creatures for real-time virtual environments. In *Proceedings of SIGGRAPH 95*.
5. Brooks, R. A. (1990). Elephants don't play chess. In Pattie Maes, ed., *Designing autonomous agents*. Cambridge, MA: MIT Press, 1990.
6. Brooks, R. A. (1997). From earwigs to humans. *Robotics and Autonomous Systems*, 20, (2-4), 291-304.
7. Bruner, J. (1990). *Actual minds, possible worlds*. MA: Harvard University Press, 1990.
8. Bruner, J. (1990). *Acts of meaning*. Cambridge, MA: Harvard University Press, 1990.
9. Dennett, D. (1987). *The intentional stance*. MIT Press, Cambridge, MA, 1987.
10. Foner, L. (1993). What's an agent, anyway? URL: foner.www.media.mit.edu/people/foner/Julia/Julia.html Published in a revised version in *The Proceedings of the First International Conference on Autonomous Agents* (AA '97).
11. Goffman, E. (1961). *Asylums: Essays on the social situation of mental patients and other inmates*. Garden City, NY: Anchor Books, 1961.
12. Janet, P. (1889). *L'Automatisme psychologique: Essai de psychologie experimentale sur les formes inferieures de l'activite humaine*. Paris: Ancienne Librairie Germer Bailliere et Cie, 1889. Ed. Felix Alcan.
13. Laing, R. D. (1960). *The divided self: An existential study in sanity and madness*. Middlesex, UK: Penguin Books, Ltd., 1960.
14. Laing, R.D. & Esterson, A. (1970). *Sanity, madness, and the family*. Middlesex, UK: Penguin Books, Ltd., 1970.
15. Loyall, A. B. (1997). *Believable agents: Building interactive personalities*. PhD thesis, Carnegie Mellon University, Pittsburgh, CMU-CS-97-123.
16. Loyall, A. B. & Bates, J. Hap: A reactive, adaptive architecture for agents. Carnegie Mellon University, Pittsburgh, *Technical Report* CMU-CS, 91-147.
17. Mateas, M. (2000). Expressive AI. *SIGGRAPH 2000 Electronic Art and Animation Catalog, 2000*.
18. Penny, S. (1997). Embodied cultural agents at the intersection of robotics, cognitive science, and interactive art. In Dautenhahn Kerstin, ed., *Socially intelligent agents: Papers from the 1997 fall symposium*, AAI Press. Menlo Park, CA, 103-105, A 1997.
19. Perlin, K. & Goldberg, A. (1996). Improv: A system for scripting interactive actors in virtual worlds. *Computer Graphics* 29, (3).
20. Reilly, W.S.N. (1996). *Believable social and emotional agents*. PhD thesis, Carnegie Mellon Univeristy, CMU-CS-96-138.
21. Reynolds, C. (1999). Steering behaviors for autonomous characters. In *1999 Game Developers Conference*. San Jose, CA, March 1999.
22. Robear, Jr, J.W. (1991). Reality check. In John G. H. Oakes, ed., *In the realms of the unreal: "Insane" writings*, 18-19. New York; Four Walls Eight Windows, 1991.
23. Ronell, A. (1989). *The telephone book: Technology - schizophrenia - electric speech*. Lincoln: University of Nebraska Press, 1989.
24. Sack, W. Stories & Social Networks (1999). 1999 *AAAI Symposium on Narrative Intelligence*. Menlo Park, CA: AAAI Press, 1999.
25. Sengers, P. (1998). *Anti-boxology: Agent design in cultural context*. PhD thesis, Carnegie Mellon University Department of Computer Science and Program in Literary and Cultural Theory, Pittsburgh, PA, 1998.
26. Sengers, P. (2000). Narrative intelligence. In Kerstin Dautenhahn, ed., Human cognition and social agent technology, advances in consciousness. John Benjamins Publishing Co, Amsterdam, 2000.
27. Smithers, T. (1992). Taking eliminative materialism seriously: A methodology for autonomous systems research. In Francisco J. Varela and Paul Bourgine, eds., *Towards a practice of autonomous systems: Proceedings of the First European Conference on Artificial Life*, 31-47. Cambridge, MA: MIT Press, 1992.
28. Steels, L. (1994). The artificial life roots of artificial intelligence. *Artificial Life*, 1,(1-2), 75-110.
29. Wavish, P. & Graham, M. (1996). A situated action approach to implementing characters in computer games. *AAI*, 10.

77