Expressive AI

Michael Mateas

Computer Science Department
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
michaelm@cs.cmu.edu

Introduction

The field of Artificial Intelligence (AI) has produced a rich set of technical practices and interpretive conventions for building machines whose behavior can be narrated as intelligent activity. Artists have begun to incorporate AI practices into cultural production, that is, into the production of artifacts and experiences that function as art within the cultural field. In this paper I describe my own practice of AI-based cultural production: expressive AI. I will attempt to provide a preliminary understanding of this practice by both situating expressive AI with respect to other discourses on AI and by working inductively from my own AI-based art work. I will first provide a brief description of three of my AI-based art pieces. These will serve as concrete examples to ground the rest of the discussion. I will then describe the expressive AI practice by first situating it with respect to the GOFAI/interactionist AI debate, then by describing the central organizing metaphors of authorial and interpretive affordance, and finally by providing a preliminary set of desiderata for expressive AI practice.

Three AI-based Artworks

This section describes three of my AI-based artworks. In these brief descriptions, I've combined a discussion of both the concept of the piece and the technical implementation. Both artists and AI researchers are likely to find these hybrid descriptions unsatisfying. However, these hybrid descriptions are necessary in order to ground the discussion of the practice of expressive AI.

Subjective Avatars

The goal of the Oz project (Bates, 1992) at CMU is to build dramatically interesting virtual worlds inhabited by believable agents - autonomous characters exhibiting rich personalities, emotions and social interactions. In many of these worlds, the player is herself a character in the story, experiencing the world from a first person perspective. Typically, the player's representation within the world - her avatar - is passive. The avatar performs actions as fully specified by the player and reports events (by, for example, rendering a 3D scene or generating descriptive text) in a pseudo-objective manner (*pseudo*-objective because any description encodes the bias of the world author). An alternative is a subjective avatar (Mateas 1997a): an avatar with autonomous interpretations of the world.

Why Subjective Avatars? I want the user to step into the shoes of a character, experiencing a story from this new perspective. In this manner the user gains an empathic understanding of a character by *being* this character. In non-interactive drama (movies, theater), an audience is able to gain insights into the subjective experience of characters precisely because the experience is non-interactive; the characters in the drama make decisions different from those that audience members might make. In an interactive story, how will a user gain insight into the character she is playing when she is controlling this character's actions? If she were to immediately begin acting out of character, she will derail the story, effectively preventing any insight. With a subjective avatar, the hope is that if the user's avatar filters and interprets the world in a manner consistent with the character, the user will begin to feel like their character, gaining a deeper understanding of the message the author wants to convey. The avatar becomes an additional artistic resource for authorial expression.

I've experimented with subjective avatars within the Oz text-based world. The text-based world accepts commands from the user and presents the world to the user in a manner similar to text-based adventure games.

Subjective State. In order for the avatar to provide a subjective interpretation for the player, it responds to activity in the world by maintaining subjective state. Currently, the avatar's subjective state consists of emotional state (emotional responses to events) and story context.

To maintain emotional state, I make use of Em (Neal Reilly, 1996), the Oz model of emotion. Em is integrated with Hap (Loyall and Bates, 1991), a reactive-planning language specifically designed for writing characters. In Em, emotions are generated primarily in response to goal processing events and attitudes. In order for the avatar to have goal processing emotions, it must be processing some goals. Since the avatar doesn't directly take action on its own, its goals are all passive. Passive goals wait for some event to occur in the world in order to succeed or fail.

In addition to emotion processing, the avatar keeps track of where it is in the story. This is done to organize the avatar's goals and simplify the writing of behaviors. At different points in the story experience, the same event may cause different reactions in the avatar (or no reaction).

Narrative Effects. Once the avatar is maintaining a subjective state, it must express this state in such a way as to affect the user's experience. The primary effect I've experimented with is manipulating sensory descriptions. Sensory manipulations are implemented as a set of Hap behaviors which render descriptions of events as a function of the subjective state. For example, imagine that the player-character (the character controlled by the human user) is afraid of a character named Barry. Barry, a manager in a fast food restaurant, is about to chew out the player. Without the subjective avatar, this would be rendered as follows in the Oz text-based world: "Barry is speaking to you. Barry's voice says 'wait a minute there, buster.' Barry goes to the counter area. Barry is no longer in the window area." The subjective avatar I've implemented for this world would render this exchange as follows: "With a vindictive gleam in his eye, Barry snaps 'Wait a minute there, buster.' Barry marches toward you from the drive-up window station." This description is generated by a narrative rule that matches on the current subjective state of the avatar (in this case, fear), and the current activity in the world. The important thing to note is that the same "objective" events in the world (Barry saying "wait a minute there, buster" and walking toward the player) would be rendered differently if the avatar felt differently (for example, as a result of previous events in the experience).

Subjective Avatar as Expressive Resource. A subjective avatar is like an inverse user model. A user model watches a user's actions so as to learn a model of the user. A subjective avatar, on the other hand, has an author given model of a character. The avatar actively manipulates a user's experience so as to try and make the user feel the same way as the character. The avatar thus becomes an active expressive resource available to dramatic world authors.

Office Plant #1

Walk into a typical, high tech office environment, and, among the snaking network wires, glowing monitors, and clicking keyboards, you are likely to see a plant. In this cyborg environment, the silent presence of the plant fills an emotional niche. Unfortunately, this plant is often dying; it is not adapted to the fluorescent lighting, lack of water, and climate controlled air of the office. Office Plant #1 (Boehlen and Mateas 1998) is an exploration of a technological object, adapted to the office ecology, which fills the same social and emotional niche as a plant. Office Plant #1 (OP#1) employs text classification techniques to monitor its owner's email activity. Its robotic body, reminiscent of a plant in form, responds in slow, rhythmic movements to express a mood generated by the monitored activity. In addition, low, quiet, ambient sound is generated; the combination of slow movement and ambient sound thus produces a sense of presence, responsive to the changing activity of the office environment. OP#1 is a new instantiation of the notion of *intimate technology*, that is, a technology which addresses human needs and desires as opposed to a technology which meets exclusively functional task specifications.

Comparable in size to a generic office plant (10x10x33 inches), OP#1 consists of a large bulb surrounded by metal fronds mounted on a base. The bulb, a hammered aluminum sphere, can open and close. Mounted on a stem, it can also rise above the fronds and remain in any intermediate position. The fronds, made of copper wire, sway slowly, moving individually or in synchrony. In addition to physical movement, OP#1 has a voice; it produces sound using a speaker housed in the bulb. These sounds provide the plant with a background presence. The force-delivering stepper motors are concealed in the lower part of the plant, discernible, though, through semitransparent plexiglas. The window in the bottom of the base would promise to reveal the inner workings of the plant, but shows, instead, a scene composed of rocks, sand and moving counterweights: the *datarium*. The datarium is the equivalent of a vivarium. In the datarium, however, the only life forms are data driven lead counterweights moving in and out of the rock and sand garden.

OP#1 is an experiment in building a companion agent, an agent that is always present, monitoring and commenting on user activity. As a constant companion, OP#1's actions must be subtle; an overactive agent would quickly becoming irritating to a user. OP#1's design attempts to maintain an air of mystery, providing a recognizable physical manifestation of a user's email activity, but not by means of a simple one-to-one mapping. OP#1 should provide the user with an opportunity for *contemplative entertainment*, opening a window onto the pattern of a user's day.

OP#1's primary view of user activity is via their email. All incoming email is assigned labels which correspond to the social and emotional role of the message, such as FYI, intimate, chatty, request, etc. Any one email may be assigned several labels. Categorization is performed by means of Naïve Bayes and K-nearest neighbor text classification (Mitchell, 1997). Naïve Bayes classifications are made by applying Bayes law to the conditional probabilities of word occurrence given a document class and the prior probabilities of document classes. The prior terms are obtained by observing frequencies in labeled training data (an offline learning step). K-nearest neighbor classifications are found by returning the majority label among the k-nearest neighbors of the query document in the document space.

The plant's behavior is controlled by a Fuzzy Cognitive Map (FCM) (Kosko, 1997). In an FCM, nodes representing actions and variables (states of the world) are connected in a network structure (reminiscent of a neural network). At any point in time, the total state of the system is defined by the vector of node values. The action associated with the action node with the highest value is executed at each point in time. The values of nodes change over time as each node exerts positive and negative influence (depending on connection weights) on the nodes it is connected to. As email is classified, activation energy is given to appropriate nodes in the network, priming OP#1's dynamics.

OP#1 is a collaboration with roboticist and artist Marc Boehlen.

Terminal Time

Terminal Time (Domike, Mateas, and Vanouse 1998, Mateas, Vanouse, and Domike 1999a) is a machine that constructs ideologically-biased documentary histories in response to audience feedback. Terminal Time is a cinematic experience, designed for projection on a large screen in a movie theater setting. At the beginning of the show, and at several points during the show, the audience responds to multiple choice questions reminiscent of marketing polls. Below is an example question.

> Which of these phrases do you feel best represents you:
>
> A. Life was better in the time of my grandparents.
> B. Life is good and keeps getting better every day.

The audience selects answers to these questions via an applause meter – the answer generating the most applause wins. The answers to these questions allow the computer program to create historical narratives that attempt to mirror and often exaggerate the audience's biases and desires. By exaggerating the ideological position implied in the audience's answers, Terminal Time produces not the history that they want, but the history that they deserve.

Critique of Traditional Historical Narratives. Terminal Time is an exploration and critique of familiar authoritarian narratives of history. Representation is at the heart of this endeavor. The mission is to dramatize to the viewing public that the truth of history is not simple and linear. Although there are undeniable historical facts, perspective is a critical element of historical understanding. By creating fact-based histories, clearly driven by point of view, the project reveals the constructed nature of all historical representation, in particular the popular genre of the television history documentary.

Representation in Terminal Time. Terminal Time represents ideological bias using a goal-tree formulation of ideology similar to Carbonell's (Carbonell, 1979). The goal tree is modified as the audience answers the polling questions. Pursuit of goals in the goal tree causes the system to search its knowledge base of historical episodes, looking for episodes which can be slanted to support the current ideological bias. In addition to historical episodes, the knowledge base also contains rhetorical devices which are used to connect episodes together to produce rhetorical flow. For example, the sentence "Yet progress doesn't always yield satisfaction" can be used to connect several episodes describing the positive effects of technological progress and several episodes describing social or environmental problems arising from technological progress. Associated with the English sentence is a formal representation constraining the meanings that episodes before and after the rhetorical device can have. Finally, Terminal Time has a media database of video clips, still images, and sounds. Each of these media elements is represented in a searchable index. Once a narrative track has been generated, Terminal Time uses the index to select media elements consistent with the narrative track.

Terminal Time is a collaboration with interactive media artist Paul Vanouse and documentary filmmaker Steffi Domike.

These three AI-based pieces provide a concrete ground for discussing expressive AI practice. They will be used as examples throughout the rest of this paper.

The GOFAI/Interactionist AI Debate

In recent years, discourse about AI's high-level research agenda has been structured as a debate between symbolist, or Good Old Fashioned AI (GOFAI), and behavioral, or interactionist AI. The GOFAI/interactionist distinction has shaped discourse both within AI and cognitive science (Brooks 1990, 1991, CogSci 1993), in cultural theoretic studies of AI (Adam 1998), and in hybrid practice combining AI and cultural theory (Agre 1997, Sengers 1998, Varela, Thompson and Rosch, 1991). This debate has shaped much contemporary practice combining AI and cultural production, with practitioners commonly aligning themselves with the interactionist camp. Because of this connection with cultural practice, it will be useful to position expressive AI relative to this debate. In this section I will briefly describe the GOFAI/interactionist debate, and diagnose why it is that contemporary cultural practitioners would find the interactionist position particularly compelling. Then I will describe how the goals of expressive AI as a practice are distinct from the goals of both the GOFAI and interactionist agendas.

Characterizing GOFAI and Interactionist AI

GOFAI is characterized by its concern with symbolic manipulation and problem solving (Brooks, 1991). A firm distinction is drawn between mental processes happening "inside" the mind and activities in the world happening "outside" the mind (Agre, 1997). GOFAI's research program is concerned with developing the theories and engineering practices necessary to build minds that exhibit intelligence. Such systems are commonly built by expressing domain knowledge in symbolic structures and specifying rules and processes that manipulate these structures. Intelligence is considered to be a property that inheres in the symbolic manipulation happening "inside" the mind. This intelligence is exhibited by demonstrating the program's ability to solve problems.

Where GOFAI concerns itself with mental functions such as planning and problem solving, interactionist AI is concerned with embodied agents interacting in a world (physical or virtual) (Brooks, 1991 and Agre, 1997). Rather than solving complex symbolic problems, such agents are engaged in a moment-by-moment dynamic pattern of interaction with the world. Often there is no explicit representation of the “knowledge” needed to engage in these interactions. Rather, the interactions emerge from the dynamic regularities of the world and the reactive processes of the agent. As opposed to GOFAI, which focuses on internal mental processing, interactionist AI assumes that having a body which is embedded in a concrete situation is essential for intelligence. It is the body that defines many of the interaction patterns between the agent and its environment.

The distinctions between the kinds of systems built by GOFAI and interactionist AI researchers is summarized in table 1.

Table 1. Contrasting properties of GOFAI and interactionist AI systems

GOFAI	Interactionist AI
Narrow/deep	Broad/shallow
Generality	Fits an environment
Disembodied	Embodied and situated
Semantic symbols	State dispersed and uninterpreted
Sense-plan-act	Reactive

GOFAI systems often attempt to *deeply* model a *narrow*, isolated mental capability (e.g. reasoning, memory, language use, etc.). These mental components duplicate the capabilities of high-level human reasoning in abstract, simplified environments. In contrast, interactionist AI systems exhibit the savvy of insects in complex environments. Interactionist systems have a *broad* range of *shallow* sensory, decision and action capabilities rather than a single, *narrow*, *deeply* modeled capability.

GOFAI seeks general solutions; *the* theory of language understanding, *the* theory of planning, etc. Interactionist AI starts with the assumption that there is a complex "fit" between an agent and its environment; there may not be generic solutions for all environments (just as many animals don't function well when removed from their environment).

GOFAI divorces mental capabilities from a body; the interface between mind and body is not commonly addressed. Interactionist AI assumes that having a body which is embedded in a concrete situation is essential for intelligence. Thus, interactionists don't buy into the Cartesian split. For them, it is the body that defines many of the interaction patterns between the agent and its environment.

Because of AI's historical affinity with symbolic logic, many GOFAI systems utilize semantic symbols - that is, pieces of composable syntax which make one-to-one reference to objects and relationships in the world. The state of the world within which the mind operates is represented by a constellation of such symbols. Interactionist AI, because of it's concern with environmental coupling, eschews complex symbolic representations; building representations of the environment and keeping them up-to-date is notoriously difficult (e.g. the frame and symbol grounding problems). Some researchers, such as Brooks (Brooks 1990, Brooks 1991), maintain the extreme position that *no* symbolic representations should be used (though all these systems employ state - one can get into nasty arguments about what, precisely, constitutes a symbol).

In GOFAI systems, agents tend to operate according to the sense-plan-act cycle. During sensing, the symbolic representation of the state of the world is updated by making inferences from sense information. The agent then constructs a plan to accomplish its current goal in the symbolically represented world by composing a set of operators (primitive operations the agent can perform). Finally, the plan is executed. After the plan completes (or is interrupted because of some unplanned-for contingency), the cycle repeats. Rather than employing the sense-plan-act cycle, interactionist systems are reactive. They are composed of bundles of behaviors, each of which describes some simple action or sequence of actions. Each behavior is appropriate under some environmental and internal conditions. As these conditions constantly change, a complex pattern of behavioral activation occurs, resulting in the agent taking action.

Interactionist AI's Affinity with Cultural Theory

Interactionist AI and GOFAI are two technical research agendas within AI, each determining a collection of research problems and system-building practices. In this section I examine the cultural theoretic association between interactionist AI and contemporary artistic practice.

Cultural theory is a diverse collection of literary, historical and sociological practices concerned with understanding the metaphors and meaning systems by which culture is composed. For cultural theorists, any cultural formation can be "read" in the same manner that one might analyze a text, seeking an understanding both of the dynamic and endlessly ramifying life the formation has within culture and the ways in which the formation is a historically contingent product of a specific cultural milieu. Cultural theory undermines the distinction between "fanciful" sign systems (e.g. literature, art) which are clearly understood as contingent, social constructions, and "true" sign systems (e.g. gender definitions, perspective vision) which are generally understood as being pre-cultural (and thus existing outside of culture). Politically, cultural studies is engaged in a project of emancipation. Social inequities are supported by unexamined beliefs (that is, "truths") about the nature of humanity and the world. For example, the inferior role of women in society is generally understood within cultural studies circles as being supported by the system of enlightenment rationality (in addition to other meaning systems). By understanding the subjugating meaning system as culturally contingent, the absolute ground from which the system operates is undermined.

Cultural theory's affinity with interactionist AI is based in a critique of Enlightenment rationality. Starting with Descartes, Enlightenment thinkers developed a theory of rationality, defining thought in terms of abstract, preferably formal operations taking place in an inner mental realm divorced from the world of gross matter. This conception of intelligence, with the twist of embedding mental operations in a material base (the brain) while still maintaining a strong split between the inner mental world and the outer world, dominates the contemporary understanding of mind. In fact, this meaning system is so hegemonic as to make it difficult to conceive of any alternative. This is precisely the kind of situation cultural theorists love to interrogate; by revealing the historical and cultural relativity (and thus rendering contingent) of the meaning system, a space of alternatives is opened up. For the case of the Enlightenment conception of mind this analysis has focused on revealing the ways in which interaction with the world, and particularly the notion of an embodied actor marked with a specific racial and sexual identity, was systematically marginalized. In keeping with the political project of cultural theory, this marginalization of embodiment has been seen as a theoretical support for the white, male subjugation of women and people of color. Interactionist AI, as a technical research agenda, seems to be reaching the same conclusions as this cultural theoretic project. Some cultural theorists explicitly acknowledge this alignment (Adam 1998). One result of this is that the moral energy associated with the political component of the cultural theoretic project transfers to the technical agenda; interactionist AI is associated with freedom and human rights and GOFAI with oppression and subjugation.

Much of contemporary arts practice is no longer concerned with the modernist agenda of perfecting purely formal elements. Rather, this practice involves self-consciously questioning cultural forms, representational modes and tropes, exploring the boundaries of these forms, breaking the representation, questioning whose power is being preserved by a representational mode, and hybridizing modes in order to create new ones, all from a position of extreme cultural self-consciousness. This self-conscious concern with meaning systems makes contemporary art practice and cultural theory natural allies, with many artists being informed by and participating in cultural theoretic analysis. And through this link with cultural theory many artists inherit their attitude towards AI, aligning with interactionist AI (and bottom-up methods in general) while feeling a generalized distrust of GOFAI, often accompanied with a sense of moral outrage acquired from cultural theory's political project. Contemporary artists thus come to see *interactionist AI as peculiarly suited for cultural production*.

Interactionist AI & Cultural Production

The expressive AI project does *not* view interactionist AI as possessing a privileged role in AI-based cultural production. Before describing the expressive AI agenda, I need to first disrupt this privileged position.

Agent as metaphor. Within the AI community, the interactionist/GOFAI debate is organized around the idea of an agent. Within AI, an agent is understood as an autonomous entity existing in an environment; it is able to sense and act on this environment. Historically, interactionist AI appeared as a reaction to recurring problems appearing in GOFAI in the design of complete agents and particularly robots (Brooks, 1990, 1991). In recent years the AI research community has indeed begun converging on reactive techniques for agent design, proposing a number of reactive and hybrid (combining search and reactivity) architectures for robotic and virtual agents. However, AI-based cultural production is broader than agent design. For example, while both Subjective Avatars and Office Plant #1 can be understood as agents, Terminal Time is not an agent (at least it can't be understood as an agent without broadening the notion of agent until it is vacuous), and yet is indisputably an instance of AI-based cultural production. In fact, Terminal Time makes heavy use of GOFAI techniques. An AI-based artist aligning herself too strongly with interactionist techniques may find that all her work becomes assimilated to the metaphor of agent, thus missing out on a rich field of alternative strategies for situating AI within culture.

Cultural production vs. AI. For the artist, even more important than recognizing the way that the metaphor of agency structures the interactionist/GOFAI technical debate is recognizing that *both* interactionist AI and GOFAI share research goals which are at odds with the goals of those using AI for cultural production. Table 2 summarizes some of the differences between cultural production and traditional AI research practice.

Table 2. Contrasting goals of cultural production and AI

Cultural production	AI
Poetics	Task competence
Audience perception	Objective measurement
Specificity	Generality
Artistic abstraction	Realism

Artists are concerned with building artifacts that convey complex meanings, often layering meanings, playing with ambiguities, and exploring the liminal region between opaque mystery and interpretability. Thus the purpose of, motivation behind, or concept defining any particular AI-based artwork will be an interrelated set of concerns, perhaps not fully explicable without documenting the functioning of the piece itself. In contrast, the focus in AI is on task competence, that is, on demonstrably accomplishing a well defined task. "Demonstrably accomplishing" means being able to show, either experimentally or by means of mathematical proof, that the AI system accomplishes the task. "Well defined task" means a simple, concisely defined objective that is to be accomplished with a given set of resources, where the objective often has "practical" (i.e. economic) utility. In GOFAI, task competence has often meant competence at complex reasoning and problem solving. For interactionist AI, this has often meant moving around in complex environments without getting stepped on, falling off a ledge, or stuck behind obstacles. In describing Office Plant #1 (OP#1) to AI practitioners (and more generally, CS practitioners), I often confront this distinction between poetics and task competence. A technical researcher tends to view OP#1 as a sophisticated email indicator that would be used to indicate to the user whether they should read their mail or not. That is, OP#1 is viewed as a mechanism for facilitating the task of reading and answering email. The notion that OP#1 is really about creating a presence whose behavior should correlate with email activity while maintaining a sense of mystery, and whose "function" is to open a contemplative window onto a "user's" daily activity, is only communicated to a technical practitioner with some difficulty.

The success of an AI-based artwork is determined by audience perception. If the audience is able to participate in the poetics defined by the artist, that is, engage in an interpretive process envisioned by the artist, then the piece is successful. AI tries to measure success objectively. How many problems could the program solve? How long did the robot run around before it got into trouble? How similar is the system's solution to a human's solution? The artist is concerned with the subjective experience of the audience, where the AI researcher strives to eliminate any reference to human perception of their artifact. All three example AI-based artworks described above are intimately concerned with audience experience. Subjective Avatars structures a participant's experience so as to help her experience a virtual world from an alien subjective viewpoint. OP#1 creates a variable sculptural presence reflecting its owner's daily activity. Terminal Time makes visible ideological bias in the construction of history by generating biased histories in response to audience feedback. There is no audience-free vantage point from which to consider these systems.

Artists build specific works. Each piece is crafted so as to establish a specific poetics, so as to engage the audience in specific processes of interpretation. The artist explores meaning-making from the vantage point of his or her particular cultural situation. AI, like most sciences, tries to create general and universal knowledge. Even interactionist AI, while stressing the importance of an agent's fit to its environment, seeks general principles by which to describe agent/environment interactions. Where AI conceives of itself as searching for timeless truths, artists participate in the highly contingent meaning systems of a particular cultural milieu. Even those AI practitioners engaged in the engineering task of building "smarter" gizmos here and now, and who would probably demure from the "timeless truth" characterization of AI practice, are still committed to building generally applicable engineering tools. Subjective Avatars provides an example of expressive AI's focus on specificity. The characters in Subjective Avatars were built using Hap, a language designed to facilitate the crafting of specific, unique characters (Loyall and Bates 1991). This is in contrast to both ALife and top-down approaches to character which attempt to define universal character frameworks in which specific characters are "tuned-in" by adjusting parameters in the model (Mateas 1997b).

Finally, artists engage in abstraction. That is, they are not so much concerned with building exact replicas of parts of the world (mimesis), as with creating meaning systems that make reference to various aspects of the lifeworld (the amalgam of the physical world plus culture). On the other hand, much of AI research is motivated by realism. A GOFAI researcher may claim that their program solves a problem the way human minds really solve the problem; an interactionist AI researcher may claim that their agent *is* a living creature, in that it captures the same environment/agent interactions as an animal. The first time I presented Terminal Time to a technical audience, there were several questions about whether I was modeling the way that real historians work. The implicit assumption was that the value of such a system lies in its veridical model of human behavior. In fact, the architectural structure of Terminal Time is part of the concept of the piece, not as a realist portrait of human behavior, but rather as a caricature of certain institutionalized processes of documentary film making.

Artistic practice transforms AI. Artistic practice is potentially concerned with a broader set of issues than the issues of agency that structure the technical interactionist/GOFAI debate. Artistic practice also operates from a different set of goals and assumptions than those shared by both interactionist and GOFAI researchers. Thus, despite the affinity between cultural theoretic critiques of Enlightenment rationality and the technical project of interactionist AI, we should be wary of any position, implicit or explicit, holding that some particular technical school of thought within AI is particularly suited to artistic practice. AI-based art is not a subfield of AI, nor affiliated with any particular technical school within AI, nor an application of AI. Rather it is a stance or viewpoint from which all of AI is reconstructed. When artistic practice and AI research combine, it results in a new interdiscipline, one I term expressive AI.

Expressive AI

AI has traditionally been engaged in the study of the possibilities and limitations inherent in the physical realization of intelligence (Agre, 1997). The focus has been on understanding AI systems as independent entities, studying the patterns of computation and interactions with the world that the system exhibits in response to being given specific problems to solve or tasks to perform. Both GOFAI and interactionist AI reify the notion of intelligence. That is, intelligence is viewed as an independently existing entity with certain essential properties. GOFAI assumes that intelligence is a property of symbolic manipulation systems. Interactionist AI assumes that intelligence is a property of embodied interaction with a world. Both are concerned with building something that *is* intelligent; that unambiguously exhibits the essential properties of intelligence.

In expressive AI the focus turns to *authorship*. The AI system becomes an artifact built by authors in order to communicate a constellation of ideas and experiences to an audience. If GOFAI builds brains in vats, and interactionist AI builds embodied insects, then expressive AI builds *cultural artifacts*. The concern is not with building something that *is* intelligent independent of any observer and their cultural context. Rather, the concern is with building an artifact that *seems* intelligent, that participates in a specific cultural context in a manner that is perceived as intelligent. Expressive AI views a system as a performance. Within a performative space the system expresses the author’s ideas. The system is both a messenger for and a message from the author.

Metaphors Structuring AI-based Artwork

The concept of an AI system as communication and performance is depicted in figure 1.

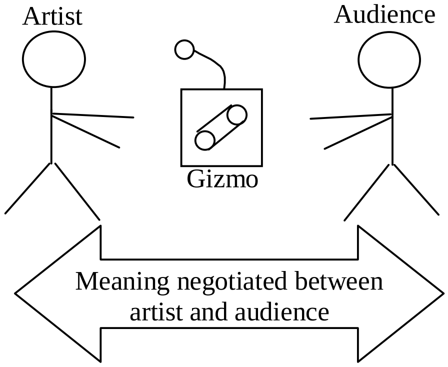

Fig. 1. The conversation model of meaning making

The AI system (here labeled "gizmo") mediates between artist and audience. The gizmo structures the context within which the artist and audience negotiate meaning. The artist attempts to influence this negotiation by structuring the interpretive affordances of the gizmo, that is, by providing the audience with the resources necessary to make up a story about what the gizmo is doing and what meanings the author may have intended to communicate. This relationship between gizmo, artist, and audience is the conversation metaphor, artistic practice conceived of as a conversation between artist and audience mediated by the art "object" (the object can be something non-concrete, such as a performance).

The conversation metaphor is an example of what Agre (Agre 1997) calls a theory-constitutive metaphor. Such a metaphor structures the theories and practices of a field. Every such metaphor has a center and a margin. The center is the set of issues brought into focus by the metaphor, those issues which will be considered primary in the practice structured by the metaphor. The margin is the set of issues made peripheral by the metaphor, those issues which will only be a secondary part of the practice, if considered at all. The practice may even assume that the margin will "take care of itself" in the process of focusing on the center.

The center of the conversation metaphor is the relationship between two subjects, the artist and the audience. A practice structured by this metaphor will focus on the negotiation of meaning between these two subjects. The margin is the internal structure of the gizmo itself. The conversation metaphor interprets the internal structure of the gizmo as an accidental byproduct of a focus on negotiated meaning; the structure "takes care of itself" in the process of focusing on the negotiation of meaning between artist and audience.

The central and marginal concerns of the conversation metaphor reverse those found in AI research practice.

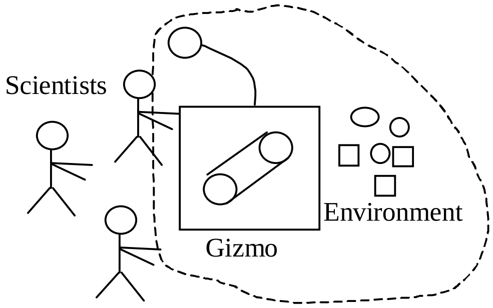

Fig. 2. The construction model of AI research

AI research practice proceeds by means of the construction metaphor. The gizmo (in GOFAI practice) or the gizmo + environment (in interactionist AI practice) is considered as a system complete unto itself, about which statements can be made without reference to the either the system builders or interpreters as subjects. Instead, system construction and interpretation is rendered as an objective process; construction is conditioned only by engineering concerns, and interpretation only by the requirements of empirical investigation. The active process of meaning making engaged in by a subject is marginalized.

Expressive AI simultaneously focuses on the negotiation of meaning *and* the internal structure of the AI system. These two apparently disparate views are unified by thinking in terms of affordances: negotiation of meaning is conditioned by interpretive affordances and the internal structure of the AI system is conditioned by authorial affordances. Before describing interpretative and authorial affordance, it is useful to first define the more general concept of affordance.

The notion of affordance was first suggested by Gibson (Gibson 1977, Gibson 1979) as a way to understand perception and was later re-articulated by Norman (Norman 1988) in the field of interface design. Affordances refer to the perceived properties of things, particularly those properties that suggest actions that can be taken with the thing. Affordances are the opportunities for action made available by an object. But affordance is even stronger than implied by the phrase "made available"; in order for an object to be said to afford a certain action, the object must in some sense "cry out" for the action to be taken. There should be a naturalness to the afforded action that makes it the obvious thing to do. For example, the handle on a teapot affords picking up the teapot with your hand. The handle cries out to be grasped. Affordances not only limit what actions can be taken (the negative form of constraint) but cry out to make certain actions obvious (the positive form of constraint).

Interpretive Affordance

Interpretive affordances support the interpretations an audience makes about the operations of an AI system. In the conversation model of negotiated meaning, it is the interpretive affordances which condition the meanings that can be negotiated between artist and audience. Interpretive affordances provide resources both for narrating the operation of the system, and additionally, in the case of AI-based *interactive* art, for supporting intentions for actions that an audience may take with the system.

Agre (Agre 1997) describes how AI technical practice provides narrative affordances which support AI researchers in creating stories describing the system's operation. Different practices (e.g. GOFAI or interactionist AI) provide different affordances for narrating system behavior. However, in typical AI research practice, these affordances are not consciously manipulated. Rather, they serve as part of the unconscious background of the engineering practice; they co-evolve with the technical practice as a silent but necessary partner in the research. Expressive AI practitioners think explicitly about how to provide the affordances supporting the narration of system behavior. For example, Sengers (Sengers 1998) explicitly added transition behaviors to behavior-based agents to support the audience's ability to narrate the agent's behavior.

For interactive art, intentional affordances support the goals an audience can form with respect to the artwork. The audience should be able to take an action and understand how the artwork is responding to this action. This doesn't mean that the artwork must provide simple one-to-one responses to the audience's actions. Such simple one-to-one responses would be uninteresting; rather, the poetics of the piece will most likely avoid commonly used tropes while exploring ambiguities, surprise, and mystery. But the audience should be able to understand that the system is responding to them, even if the response is unexpected or ambiguous. The audience should be able to tell some kind of unfolding story about their interaction with the work. Both extremes, simple stereotyped responses to audience interaction making use of well-known tropes, and opaque incoherence with no determinable relationship between interaction and the response of the art work, should be avoided.

A concern with interpretive affordance will be familiar to artists; negotiating meaning between artist and audience is central to artistic practice. Expressive AI adopts this concern within the context of AI-based art. But expressive AI also adopts a concern for the artifact from AI research practice.

Authorial Affordance

The authorial affordances of an AI architecture are the "hooks" that an architecture provides for an artist to inscribe their authorial intention on the machine. Different AI architectures provide different relationships between authorial control and the combinatorial possibilities offered by computation. Expressive AI engages in a sustained inquiry into these authorial affordances, crafting specific architectures that afford appropriate authorial control for specific art works.

This concern with the machine itself will be familiar to AI research practitioners (both GOFAI and interactionist). However, AI research practice tends to focus on individual architectures, not on the human authorship supported by the architecture nor on understanding the differences between architectures. AI research practice downplays the role of human authorship within the system because this authorship disrupts the story of the system as an autonomously intelligent entity. Rather, the focus is on the architecture itself, independent of any "content", and generally independent of any discussion of any other architecture. Expressive AI simultaneously adopts and transforms this concern with the machine.

A focus on the machine is alien to current electronic media practice. In keeping with the conversation metaphor of meaning making, the internal structure of the machine is generally marginalized. The machine itself is considered a hack, an accidental byproduct of the artist's engagement with the concept of the piece. In the documentation of electronic media works, the internal structure of the machine is almost systematically effaced. When the structure is discussed, it is usually described at only the highest-level, using hype-ridden terminology and wishful component naming (e.g. "meaning generator", "emotion detector"). At its best, such discursive practice is a spoof of similar practice within AI research, and may also provide part of the context within which the artist wishes her work to be interpreted. At its worst, such practice is a form of obfuscation, perhaps masking a gap between intention and accomplishment, the fact that the machine does not actually do what is indicated in the concept of the piece.

Why would an artist want to concern herself with authorial affordance, with the structural properties of the machine itself? Because such a concern allows an artist to explore expressive possibilities that can only be opened by a simultaneous inquiry into interpretive affordance and the structural possibilities of the machine. An artist engaging in expressive AI practice *will be able to build works with a depth, a richness, a sophistication that can't be achieved without this simultaneous focus on meaning making and machine structure.*

Combining Interpretive and Architectural Concerns

The splitting of AI-based art practice into interpretive and authorial concerns is for heuristic purposes only, as a way to understand how expressive AI borrows from both art practice and AI research practice. Expressive AI practice combines these two concerns into a dialectically related whole; the concerns mutually inform each other. The "interface" is not separated from the "architecture." In a process of total design a tight relationship is maintained between the sensory experience of the audience and the architecture of the system. The architecture is crafted in such a way as to enable just those authorial affordances that allow the artists to manipulate the interpretive affordances dictated by the concept of the piece. At the same time, the architectural explorations suggest new ways to manipulate the interpretive affordances, thus suggesting new conceptual opportunities.

The AI-based artist should avoid architectural elaborations which are not visible to the audience. However, this admonition should not be read too narrowly. The architecture itself may be part of the concept of the piece, part of the larger interpretive context of people theorizing about the piece. For example, one can imagine building a machine like Terminal Time in which some small finite collection of historical narratives have been prewritten. The narrative played is determined by a hard-coded selection mechanism keyed off the audience polls. For any one audience, the sensory experience of this piece would be indistinguishable from Terminal Time. However, at a conceptual level, this piece would be much weaker than Terminal Time. A Terminal Time audience is manipulating a *procedural process* which is a caricature of ideological bias and of institutionalized documentary filmmaking. The operationalization of ideology is critical to the concept of the piece, both for audiences and for artists and critics who wish to theorize the piece.

Why Use AI in Cultural Production?

At this point the practice of expressive AI has been described as one combining both a focus on meaning-making and the authorial affordances of AI architectures. However, this begs the question of why an artist would want to use AI in cultural production at all. Here I enumerate some of reasons *I* engage in AI-based art practice.

Support sophisticated modes of interaction. AI-based interactive art can respond to audience interaction with a sophistication that is not possible without AI techniques. For example, with Subjective Avatars the audience manipulation of the avatar causes a complex pattern of processing to occur in a behavior model of a specific personality (whatever role the audience is "playing") resulting in an active manipulation of the audience's experience of the world.

Procedural portraits of human meaning-making. AI techniques support the construction of procedural portraits of human meaning-making. A procedural portrait is a representation of some human cultural process. For example, Terminal Time is a procedural portrait of the ideologically-biased construction of mainstream historical documentaries.

Actively participate in the realm of human meaning. AI-based art can directly observe and act on activities laden with human meaning. For example, Office Plant #1 is able to react to the social and emotional content of email; this requires that it have some window on the human interpretation of email.

Tap into rich history of narrative affordance. As discussed in the section on interpretive affordance, any interactive artwork must provide the resources for an audience to interpret the activities of the artwork. The technical practice of AI has a rich history of constructing machines with narrative affordances (albeit the existence of these affordances are usually not acknowledge). This practice provides a fertile field for building machines that afford complex interpretation.

Expressive AI Desiderata

Now that the practice of expressive AI has been given an abstract description, this section provides a tentative list of desiderata.

Expressive AI is not "mere application." Expressive AI is not an application area of AI. Applications are understood as the use of off-the-self techniques which are unproblematically appropriated to some concrete task. AI applications do not question the deep technical and philosophical assumptions that underlie AI practice. Expressive AI, on the other hand, changes AI practice by simultaneously exploring interpretive and authorial affordances. Expressive AI is not a technical research program calling for the overthrow of GOFAI or interactionist AI. Nor does it single out a particular technical tradition as being peculiarly suited for artistic expression. For example, subjective avatars draw from interactionist AI, Office Plant #1 draws from statistical AI, and Terminal Time draws from GOFAI. Rather, expressive AI is a stance or viewpoint from which AI techniques can be rethought and transformed. New avenues for exploration are opened up; research values are changed.

Build microworlds with human significance. Building microworlds was an AI approach popular in the 1970s. The idea was to build simple, constrained, artificial worlds in which an AI system could exhibit its competence. The hope was that it would be possible to slowly scale up from systems that exhibit competence in a microworld to systems exhibiting competence in the real world. The microworld research agenda has been widely criticized (e.g. Dreyfus 1999); it did not prove possible to scale systems up from microworlds. However, the microworld concept can be useful in expressive AI. An AI-based art piece may be a microworld with human significance. The "micro" nature of the world makes certain AI techniques tractable. As long as the microworld has some cultural interest, the system still functions as an artwork. This is simply the recognition that an artwork is not the "real world" but is rather a representational space crafted out of the world. The AI techniques used in an artwork only have to function within the specific artistic context defined by the piece. For example, in Subjective Avatars, the agents only have to operate within the specific dramatic context defined by the storyworld.

Actively reflect on affordances associated with different architectures. Expressive AI practitioners must unpack the complex relationships that exist between authorial intention and different architectures. Architectures, and the associated technical practices supporting the architecture, make available different authorial and interpretive affordances. Active reflection on the co-evolution of affordances and technical solutions is part of expressive AI considered as a design practice. By understanding these relationships, the practitioner improves her skill as an AI-based artist, becoming more able to navigate the design space of affordance plus architecture. While this reflection is similar to AI research practices, it differs in focusing explicitly on affordances, which are commonly left unarticulated in traditional AI practice.

Cultural theory and expressive AI. In the first part of this paper I took pains to undermine any claim interactionist AI might have for being peculiarly suited to artistic practice by diagnosing the link that exists between cultural theoretic critiques of Enlightenment rationality and interactionist AI. This may have left the reader with the impression that I am hostile to cultural theoretic studies of AI. This is not the case. Culture theory is extremely valuable for unpacking hidden assumptions lurking in AI practice. Understanding these assumptions allows an artist to gain a free relation to AI technology, to avoid being forced into the "natural" interpretation of the technology that has been historically constructed. It is only the implicit claim that a particular technology is suited for artistic expression that expressive AI rejects. Cultural studies of AI help a practitioner to maintain a free relation to technology, but this is a process, not an achievable end. There is no final, "perfect" AI to be found, for artistic or any other purpose.

Computer games as a high-art form. AI-based interactive art has the potential to hybridize with computer games to form a new mass-audience high-art form (Mateas 1999b). Electronic media art is already stretching the boundaries of the gallery and museum space. Perhaps, like cinema before it, electronic media art will need a new venue in order to become broadly accessible. AI-based interactive art already bears some similarity to computer games. Interactive drama is related to the already established form of the adventure game, though it differs in its focus on the first-person experience of a dramatic arc rather than goal-based puzzle solving. Office Plant #1 shares a focus on long-term engagement with virtual pets such as Dogz and Catz (Stern, Frank, and Resner, 1998), though virtual pets are intended for circumscribed, high-intensity interaction while OP#1 provides continuous, ambient commentary. These similarities hint that AI-based art could be disseminated in a manner similar to computer games, inhabiting the new cultural niche of "high-culture" interactive experiences.

I sometimes call my own practice AI-based art *and* entertainment as a way to indicate my interest in blurring the art/entertainment distinction. This distinction is really found in the culture of production, not the culture of reception. Cultural producers find it important to distinguish themselves from the "low-culture trash mongers" (if they are artists) or from the "elitists who produce only for themselves" (if they are entertainers). In the culture of reception (cultural consumers) this distinction is not sharp; it is part of a continuum ranging from "brain-dead" entertainment to "edifying" entertainment. This fluidity in the culture of reception makes the hybridization of AI-based art and computer games viable.

Conclusion

Expressive AI is a new interdiscipline of AI-based cultural production combining art practice and AI research practice. Expressive AI changes the focus from an AI system as a thing in itself (presumably demonstrating some essential feature of intelligence), to the communication between author and audience. The technical practice of building the artifact becomes one of exploring which architectures and techniques best serve as an inscription device within which the authors can express their message. Expressive AI does not single out a particular technical tradition as being peculiarly suited to culture production. Rather, expressive AI is a stance or viewpoint from which all of AI can be rethought and transformed.

References

Adam, A. 1998. *Artificial Knowing: gender and the thinking machine*. London: Routledge.

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

Bates, J. 1992. Virtual Reality, Art, and Entertainment. *Presence: The Journal of Teleoperators and Virtual Environments* 1(1): 133-138.

Boehlen, M., and Mateas, M. 1998. Office Plant #1: Intimate space and contemplative entertainment. *Leonardo*, Volume 31 Number 5: 345-348.

Brooks, R. 1991. Intelligence Without Reason, A.I. Memo 1293. Artificial Intelligence Lab. MIT.

Brooks, R. 1990. Elephants Don't Play Chess. *Robotics and Autonomous Systems* 6: 3-15.

Carbonell, J. 1979. Subjective understanding: Computer models of belief systems. Ph.D. diss., Computer Science Department, Yale University.

CogSci. 1993. Special Issue on Situated Cognition.

Cognitive Science 17 (1993)

Domike, S.; Mateas, M.; and Vanouse, P. 1998. *The recombinant history apparatus presents: Terminal Time*. Forthcoming in book from the Center for Twentieth Century Studies.

Dreyfus, H. 1999. *What Computer Still Can't Do: A Critique of Artificial Reason.* MIT Press, original edition published in 1972.

Gibson, J. 1979. *The ecological approach to human perception*. Boston: Houghton Mifflin.

Gibson, J. 1977. The theory of affordances. In R. E. Shaw & J. Bransford (Eds.), Perceiving, acting, and knowing. Hillsdale, NJ: Erlbaum Associates.

Kosko, B. 1997. *Fuzzy Engineering.* New York: Simon & Schuster, pp. 499–525.

Loyall, A. B.; and Bates, J. 1991. Hap: A Reactive, Adaptive Architecture for Agents. Technical Report CMU-CS-91-147. Department of Computer Science. Carnegie Mellon University.

Mateas, M., Vanouse, P., Domike S. 1999a. Terminal Time: An Ideologically-biased History Machine. AISB Quarterly, Special Issue on Creativity in the Arts and Sciences, Summer/Autumn 1999, No. 102, 36-43.

Mateas, M. 1999b. Not your Grandmother’s Game: AI-Based Art and Entertainment. Working notes of the AI and Computer Games Symposium, AAAI Spring Symposium Series. Menlo Park: Calif.: AAAI Press.

Mateas, M. 1997a. Computational Subjectivity in Virtual World Avatars. *Working notes of the Socially Intelligent Agents Symposium, AAAI Fall Symposium Series.* Menlo Park: Calif.: AAAI Press.

Mateas, M. 1997b. An Oz-Centric Review of Interactive Drama and Believable Agents. Technical report CMU-CS-97-156. Computer Science Department, Carnegie Mellon University.

Mitchell, T. 1997. *Machine Learning*. New York: McGraw-Hill, p. 180.

Neal Reilly, W. S. 1996. Believable Social and Emotional Agents. Ph.D. diss., School of Computer Science, Carnegie Mellon University.

Norman, D. 1988. *The Design of Everyday Things.* New York: Doubleday.

Sengers, P. 1998. Anti-Boxology: Agent Design in Cultural Context. Ph.D. diss., School of Computer Science, Carnegie Mellon University.

Stern, A.; Frank, A.; and Resner, B. 1998. Virtual Petz: A hybrid approach to creating autonomous, lifelike Dogz and Catz. In Proceedings of the Second International Conference on Autonomous Agents, 334-335. Menlo Park, Calif.: AAAI Press.

Varela, F., Thompson, E., Rosch, E. *The Embodied Mind: Cognitive Science and Human Experience*. MIT Press, seventh printing 1999.