

# Reading Hal: Representation and Artificial Intelligence

Michael Mateas  
The Georgia Institute of Technology  
michaelm@cc.gatech.edu

In this chapter I wish to focus on Hal 9000. Rather than reading Hal as a Frankensteinian cautionary tale, a representation of our disquiet over the cybernetic blurring of the human, of our fear of an evolutionary showdown with increasingly autonomous technologies, I'd like to read Hal as a representation of the goals, methodologies and dreams of the field of Artificial Intelligence (AI). As a representation, Hal, and the role he plays within *2001*, both captures preexisting intellectual currents that were already operating within the field of AI, and serves as an influential touchstone that had a profound impact on individual AI practitioners and on the aspirations of the field.

I come at this understanding of Hal from a disciplinary position that straddles the humanities, computer science, and digital art practice. While my degree is in computer science, specifically in AI, my research focus is on AI-based interactive art and entertainment. Consequently, my research agenda brings to bear new media studies and science studies, digital art practice, and technical research in AI. It is from this hybrid position, working in the context of a joint appointment in both the humanities and computer science, that I wish to read Hal as a representation of technical practice within AI.

In addition to reading Hal as a depiction of the disciplinary machinery of AI, Hal of course also functions as a character within the narrative machinery of *2001*, a

Mateas, M. 2005. Reading Hal: Representation and Artificial Intelligence. In Robert Kolker (Ed.), *Stanley Kubrick's 2001: A Space Odyssey : New Essays*, Oxford University Press, 2005 (in press).

character, as many have pointed out, with more emotional and psychological depth than any of the human characters. Once Hal is understood as a cinematic representation that simultaneously depicts specific agendas and assumptions within AI and performs an expressive function for an audience (ie. serves as a character within a story), it is a small step to consider AI systems themselves as *procedural representations* that simultaneously encode agendas and assumptions and perform for an audience. The last section of this chapter will investigate Expressive AI, that is, AI considered explicitly as a *medium*.

## **Hal and AI**

Hal was, and still is, a powerful inspiration for AI researchers. In *Hal's Legacy*<sup>1</sup>, prominent members of the AI research community describe both how Hal influenced their own work and the relationship between Hal and the current state of AI research. There have of course been many depictions of robots and intelligent computers in Sci Fi films, but few of these representations have achieved, for AI researchers, Hal's emblematic status. Unlike other Sci Fi representations of AI, Hal is special because of the way he connects to technical agendas within AI research.

Hal convincingly integrates many specific capabilities, such as computer vision, natural language processing, chess playing, etc., demonstrating the elusive generalized intelligence sought by AI researchers. Most filmic representations of AI act just like people, adding a few mechanical affectations to a clearly human performance. There is no clear relationship between these filmic representations and current lines of research in AI. Hal, on the other hand, appears as a plausible extrapolation from current lines of work, serving as a visualization for the AI community of future AI systems.

Because achieving general intelligence is difficult to turn into a pragmatic research plan, AI research tends to proceed by attacking sub-problems. The problem of creating an intelligent machine is either broken up into deep models of isolated capabilities (e.g. visually recognizing objects, creating plans of action in simplified domains), or broken up into systems that integrate a range of more shallow competencies (e.g. a robot that integrates simple sensing and planning in order to carry out a single task). In both cases the systems lack general intelligence, the ability to integrate a broad range of knowledge and physical competencies, to apply knowledge from one domain to another, to handle unexpected and new situations. AI systems only perform intelligently on a single, narrow task or within a single, simplified domain.

Hal presents to researchers a powerful cinematic representation of AI precisely because he simultaneously demonstrates general intelligence while keeping visible the AI sub-problems, roughly corresponding to different sub-fields within AI. Thus researchers can easily recognize AI specialties in Hal's individual capabilities, making Hal plausible, while seeing the individual capabilities integrated into a general intelligence, making Hal compelling. Marvin Minsky, one of the founders of AI, served as a technical consultant on the film; doubtless his contribution helped to establish the strong resonance between the depiction of Hal and sub-fields within AI, including language, common sense reasoning, computer vision, game playing, and planning and problem solving.

Language is one of the hallmarks of intelligence – natural language processing has been part of the AI research agenda since the beginning of the field. Hal demonstrates a range of natural language competencies, including understanding (making sense of sentences and conversations), generation (generating responses), speech recognition, and

speech generation. Hal is able to participate in conversations ranging from simple commands, such as Poole's commands to raise and lower his headrest and to display his parents recorded birthday greeting in his room, to complex conversations where Hal expresses inner conflicts and tells sophisticated lies. In work on natural language processing, researchers quickly discovered that generalized natural language capabilities require common sense reasoning, that is, a huge amount of knowledge about everyday objects, events and situations. This background knowledge is needed not only to disambiguate meaning through context, but also to work out the ramifications of utterances: an utterance doesn't just have a denotative meaning, but also a complex halo of connotative meanings and implications for both the speaker and listener. The common sense reasoning problem is enormous and unsolved. The problem with common sense is that it isn't really a sub-problem of the sort that AI researchers typically tackle, but rather seems to be the whole of intelligence; if you have common sense reasoning, you'd have general intelligence. For this reason, AI systems that use natural language only function within micro-domains, specific, simplified domains of expertise. For example, research into dialogue systems (systems that are able to have an extended dialogue) generally takes place in task-based domains such as travel planning<sup>2</sup>, where the system creators are able to assume that all utterances relate directly to the task at hand, and where connotative meaning is kept to a minimum.

Hal, on the other hand, demonstrates general language and common-sense reasoning capabilities. This is made plausible for an AI audience by sneaking this general competency in through the back door of an apparent micro-domain. As the shipboard computer for the Discovery, Hal's primary function is to manage the ship and participate

in the mission. Though Hal is certainly introduced as an extremely advanced AI system during the initial interview with the BBC reporter, this interview establishes Hal as a primarily functional, though advanced, onboard control system for the Discovery. As the plot progresses, Hal gradually exhibits full-blown general language competency and common-sense reasoning from within this micro-domain.

Hal's ability to see is emphasized throughout the film by frequent cuts to his camera eye and by occasionally giving the viewer a subjective view through Hal's cameras. Hal demonstrates computer vision capabilities far more sophisticated than anything we're capable of today. His vision is fully integrated with the rest of his intelligence, allowing him to, for example, talk about what he sees (integrating natural language processing and vision) or use his vision in pursuit of goals, as when he reads the astronaut's lips to stay ahead of his adversaries. Again, his visual capabilities are made plausible for an AI audience by demonstrating specific visual sub-problems. For instance, when Hal asks to see Bowman's drawings, Hal is able to recognize the objects depicted in the drawings, including the face of one of the hibernating crewmen. Object and face recognition is one of the standard well-defined sub-problems of computer vision; by giving the audience a view of the drawings through Hal's eye, the film emphasizes the specific object-recognition task Hal is engaged in. However, within this same scene, Hal moves beyond mere object recognition by commenting on Bowman's drawing style and comparing his current drawings to previous ones. By framing this discussion of style within an implicit object recognition task framework, the film presents reasoning about style and aesthetics as simple technical extensions of an understood AI research problem.

Game playing, and in particular chess, is one of the classic AI problems; early successes in chess playing were partly responsible for overly-optimistic predictions made during the 1960s and 1970s for the achievement of general machine intelligence. Because chess is considered a difficult game, something that “intelligent people” do, it was assumed that if computers could play chess, then they must also be “intelligent”. It turns out, however, that the really difficult tasks to make a computer do are generally not the tasks that humans consider difficult, such as chess playing, but everyday “easy” activities, such as using language, seeing the world and understanding what you see, common sense reasoning, and so forth. At the time that *2001* came out, AI was still in its early, optimistic phase, buoyed by successes on problems such as chess. The chess-playing scene therefore had special resonance for the AI audience – though Hal’s chess performance may have been better than 1969-era chess-playing programs, chess was a well-understood problem. AI researchers had every confidence that, in the not too distant future, there would be chess playing programs better than any human player. The chess scene thus establishes plausibility by demonstrating an easy extrapolation from the current state of a well-understood problem.

The AI sub-field of planning and problem solving is concerned with modeling goal-driven activity, that is, how intelligent systems arbitrate between multiple goals and construct and follow plans of actions to accomplish goals. Hal demonstrates goal-driven behavior in his handling of the “failure” of the AE-35 unit. After reporting a fictitious failure in the AE-35 unit, Hal expresses confusion when the diagnostic analysis reveals no failure (Hal: “Yes, it's puzzling. I don't think I've ever seen anything quite like this before.”), and suggests that the unit be replaced until it fails. Hal’s confusion about the

AE-35 unit can be read two ways. Either he is quite self-consciously lying about the AE-35 unit as part of some master plan to sever communications with mission control and lure the astronauts out into space where he can kill them, or he is genuinely confused about the AE-35, indicating an internal conflict. For an AI audience, both cases are clear instance of goal-based behavior. In the first case, Hal has a goal to eliminate the astronauts, whom he has identified as dangerous to the success of the mission, and has generated an elaborate plan to eliminate them, a plan within which he is able to improvise when the situation changes, such as when Bowman forgets his space helmet when he goes out to retrieve Poole. Hal sees that Bowman has forgotten his helmet, which enables Hal to achieve his goal of eliminating Bowman, since now all Hal has to do is refuse Bowman entry. (We can only speculate about what Hal's plan would have been had Bowman not forgotten his helmet, perhaps teleoperating a second pod to disable Bowman's pod.) In the second case, Hal's behavior can be interpreted as a goal conflict a situation in which some of Hal's actions, such as reporting the fault in the AE-35, are executed in pursuit of one goal, while other actions, such as the actions to diagnose the fault, are executed in the pursuit of a different goal, with the result that Hal's overall behavior is incoherent. The goal-based behavior evident in either reading resonates strongly for an AI audience because of the connection with the sub-field of planning and problem solving.

In addition to referencing specific sub-fields within AI, Hal also resonates with the AI audience through indirect references to the Turing Test. Alan Turing, in his seminal article on machine intelligence, sought to replace philosophical arguments about whether machines can think with an operational definition of intelligence.<sup>3</sup> In the Turing

Test, a human judge engages in typed conversation, through a terminal, with both a human and a machine that are present in another room.<sup>4</sup> The judge must determine, based on the responses to her typed queries, which is the human and which is the machine. If the judge can't tell the difference, we deem the machine "intelligent." The notion that something is intelligent if it seems intelligent, and more generally, that questions of identity (essence) should be replaced with questions about functional or behavioral equivalence, is generally accepted by AI practitioners.

In the interview with the BBC reporter, when asked if Hal has emotions, Bowman responds:

Well, he acts like he has genuine emotions. Of course he's programmed that way to make it easier for us to talk to him. But as to whether or not he has real feelings is something that I don't think anyone can truthfully answer.

This reference to behavioral equivalence immediately cues the AI audience. The issue of "genuine emotion" is has been replaced with behavioral equivalence; Hal acts like he has emotions, so he should be treated as having emotions. This move establishes a double perspective throughout the rest of the film. Whenever Hal acts in a human-like way, the audience (particularly the AI audience) simultaneously reads Hal's behavior at face value, as the behavior of a thinking, feeling, conscious being, and sees it as a consequence of entirely mechanical, comprehensible, functional processes. Bowman and Poole explicitly refer to this ambiguous double reading during their discussion of Hal's malfunction (the discussion in the pod). In the mechanical view, Hal is simply a faulty component that may have to be disconnected; this is the view unproblematically

adopted by Poole. Bowman, however, expresses concern that no 9000 series computer has ever been disconnected before and that he's not sure what Hal will think about this. The tension of this double reading peaks during Hal's final scene, as he expresses fear and pain during the disconnection of his higher brain functions ("I'm afraid, Dave. Dave, my mind is going. I can feel it."). The audience is caught between reading this as the output of a machine or the words of a being towards whom we have moral responsibility; in the context of the Turing test, both are true.

### **Classical AI**

In recent years, discourse about AI's high-level research agenda has been structured as a debate between symbolist, classical AI (sometimes called Good Old Fashioned AI or GOF AI), and behavioral, or interactionist AI. The classical/interactionist distinction has shaped discourse both within AI and cognitive science, in cultural theoretic studies of AI, and in hybrid practice combining AI and cultural theory. *2001* was released during the ascendancy of classical AI, and indeed, Hal accurately represents the vision of classical AI.<sup>5</sup>

Classical AI is characterized by its concern with symbolic manipulation and problem solving. A firm distinction is drawn between mental processes happening "inside" the mind and activities in the world happening "outside" the mind.<sup>6</sup> Classical AI's research program is concerned with developing the theories and engineering practices necessary to build minds exhibiting intelligence. Such systems are commonly built by expressing special-purpose knowledge about a specific task (such special-purpose knowledge is typically called "domain knowledge") as symbolic structures and

specifying rules and processes that manipulate these structures. Intelligence is considered to be a property that inheres in the symbol manipulation happening “inside” the mind.

This intelligence is exhibited by demonstrating the program’s ability to solve problems.

Where classical AI concerns itself with mental functions such as planning and problem solving, interactionist AI is concerned with embodied agents interacting in a physical or virtual world. 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 classical AI, which focuses on internal mental processing, interactionist AI assumes that having a body 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. For the interactionist, a body is necessary even for forming abstract concepts; abstractions are based on sensory-motor experience.

The bodiless Hal engages in the sense-plan-act cycle of classical AI. During sensing, an internal representation of the state of the world is updated by making inferences from sensory information. In addition to containing a model of the state of the world, this interior symbolic space contains the goals or intentional structure of the AI system. The system then constructs a plan of action for accomplishing goals, given the represented state of the world. Finally, the system carries out this plan in the world and begins the cycle anew. Hal’s physicality, his sensory-motor apparatus, is distributed throughout the ship in the form of cameras, microphones, and the myriad of ship systems

he can control. As a classical system, the particular configuration of Hal's sensory-motor system has no effect on his interior, mental structure. If new sensors, for example infrared cameras, or new effectors, for example a robotic arm attached to the main console, are added to Hal, it doesn't change the way Hal thinks; he merely would have new physical capabilities. For interactionist AI, the particular shape or configuration of the body strongly effects the mind; change the structure of the body and you change the structure of the mind. Hal's bodiless cognition makes him a clear example of classical AI.

Hal's radical interiority, an internal mental (symbolic) space into which neither Bowman, Poole nor we as viewers have access, is emphasized through the use of "reaction" shots focusing on one of Hal's camera eyes. Where normally a reaction shot reveals, through bodily (including facial) position and movement, a character's motivational and emotional response to a situation, letting us into a character's interior space, for Hal, who is in some sense pure mind, the reaction shots remain opaque, giving the viewer a sense of an interior they are not allowed to enter. This technique is used, with increasingly chilling effect, starting with the first hint of Hal's malfunction. During the conversation in which Hal reveals his concerns about the mission to Bowman, the frequent cuts to Hal's eye during the conversation give us a sense of depths in Hal, while keeping those depths mysterious. When Dave responds to Hal's concerns with the mild rebuke "You're working up your crew psychology report?", the camera focuses on Hal's eye and holds for a beat before Hal responds "Of course I am. Sorry about this. I know it's a bit silly." Hal then interrupts himself with the clipped "Just a moment. Just a moment." that signals the full onset of his psychosis. Similarly, during the death scene of the hibernating crewman, the camera cuts between the life support alarms and one of

Hal's expressionless camera eyes, further reinforcing a sense of complex interior machinations that remain inaccessible in the world of bodies.

The chess-playing scene further establishes Hal as an instance of classical AI. Chess, with its properties of a completely knowable world (the board), deterministic interactions (the rules), simple evaluation of success (win, lose, draw), but still offering within this simple framework a huge range of strategic and tactical options, was a popular domain for classical AI research. The simple and noise-free nature of chess effectively trivialize the sense and act portions of the sense-plan-act cycle, squarely placing the focus on internal representation and reasoning. Hal's facility with chess would have strongly resonated with 1970's era AI researchers, serving as an indicator that Hal is indeed intelligent.

Interactionist AI researchers, who see mind arising out of the behavioral details of physical interaction in the world, are more likely to resonate with Sci Fi representations that emphasize robots and androids. In the episode "Robot's Alive" of *Scientific American Frontiers* featuring interactionist AI researcher Rodney Brooks, Brooks mentions that the ultimate goal of his work is to be able to build Lieutenant Commander Data, an android in *Star Trek: The Next Generation*.<sup>7</sup> However, unlike Hal, Data fails to demonstrate general intelligence while maintaining connections to specific technical sub-problems within interactionist AI research. While Data is certainly a likeable character, and various *Star Trek* plots have explored the philosophical problems of Data's personhood, he fails to achieve the same inspiring plausibility, the magic of making general intelligence seem a natural extrapolation from current technical work. For AI

researchers, Hal remains a uniquely powerful and influential popular media representation of AI.

### **AI and Transcendence**

The environment of the Discovery is cold and antiseptic, dominated by hospital-white consoles and information displays, an environment in which the emotionless astronauts live completely scheduled lives dominated by formal procedures and routines, watched over by the infallible rationality of Hal. The Discovery is a perfect Taylorist environment, a Closed World in which all contingencies have been modeled and accounted for, at least until Hal's fatal malfunction. Hal's rationality, and his physical manifestation in the total (and totalizing) environment of the Discovery, form an odd disjunction with the film's last sequence, Bowman's transcendent rebirth as the Star Child. Apparently rationality must be defeated, Hal deactivated, before transcendence can occur. However, within the culture of AI, Hal and the birth of the Star Child are not contradictory, but are rather part of the same agenda: the transcendence of the human through the creation of thinking machines.

In *The Religion of Technology*, David Noble traces the explicitly religious, primarily Christian drive to transcend the body and the material world that operates as part of the disciplinary logic of atomic weapons science, space exploration, genetic engineering and AI. Noble argues that AI continues in the Cartesian tradition of a strong separation of mind and body, with the thinking abstract mind seen as having a direct relationship to God, and hence to truth, while the body, with its sensory animal appetites, distracts the divine, thinking mind.<sup>8</sup>

The foundational move in AI, particularly classical AI, is to view mind as an abstract process, something that is not necessarily tied to the contingencies of human brains and bodies, but can rather be abstracted and run on multiple hardware platforms, including digital computers. Minsky has described the human brain as a mere “meat machine,” and the body, that “bloody mess of organic matter,” as a “teleoperator for the brain.”<sup>9</sup> Mind is a process, a collection of functional relationships; it is only an accident of history that mental processes are implemented on the organic brains of human beings. If mind can be released from the shell of the body, running free on ever faster, more efficient hardware, it is only a matter of time before these minds achieve human-level, then superhuman intelligence.

In this AI eschatology there is an intermediate period before machine intelligence surpasses the human, a period in which human and machine intelligence work together in tightly integrated human-machine symbiosis. This is the era of the cyborg, the era in which we live now, in which the focus is on the computer as an infinitely flexible medium that extends the thinking capabilities of the human mind. J.C.R. Licklider, influential MIT psychologist and first head of the computer research program at the US government’s Advance Research Projects Agency (ARPA), established ARPA’s long-term funding of both AI and advanced human-computer interaction and communication techniques, including the development of ARPAnet, which eventually became the internet (ARPA was eventually renamed to the present-day DARPA, the Defense Advance Research Projects Agency). In 1960 he discussed the relationship between human-computer systems and the ultimate goals of AI.

Man-computer symbiosis is probably not the ultimate paradigm for complex technological systems. It seems entirely possible that, in due course, electronic or chemical ‘machines’ will outdo the human brain in most of the functions we now consider exclusively within its province. ... In short, it seems worthwhile to avoid argument with (other) enthusiasts for artificial intelligence by conceding dominance in the distant future of cerebration to machines alone. There will nevertheless be a fairly long interim during which the main intellectual advances will be made by men and computers working together in intimate association.<sup>10</sup>

Hal represents the extreme end of the era of human-computer symbiosis, a thinking tool able to function on its own, ready, and in this case, willing, to be free of its human users. For a brief period (Licklider describes this period as being somewhere in the range “15 to 400 years”), machines augment human intelligence in a symbiotic union; ultimately however, artificial intelligence exceeds human intelligence. At this stage the new machine superintelligences break out on their own evolutionary path. Like the dinosaurs, humans, in their current, messy, wet, biological form, are left far behind, an evolutionary experiment that had its day, but has been superseded by infinitely-more accomplished machine minds.

Yet personal human identity need not be lost, for the advent of superhuman machine intelligence also signals the advent of immortality. In *Mind Children*, Hans Moravec, a robotics researcher at Carnegie Mellon University who developed early, influential autonomous robots while at Stanford, describes a technical vision in which human minds are uploaded out of biological bodies into superior robotic bodies with

super-fast computer brains.<sup>11</sup> In this new substrate of silicon and steel, human minds can run much faster than on wet brains, allowing the uploaded individual to think, learn, develop and change at superhuman speeds.

Your new abilities will dictate changes in your personality. Many of the changes will result from your own deliberate tinkering with your own program. Having turned up your speed control a thousandfold, you notice that you now have hours (subjectively speaking) to respond to situations that previously required instant reactions. You have time, during the fall of a dropped object, to research the advantages and disadvantages of trying to catch it, perhaps to solve its differential equations of motion. You will have time to read and ponder an entire on-line etiquette book when you find yourself in an awkward social situation... In general, you will have time to undertake what would today count as major research efforts to solve trivial everyday problems.<sup>12</sup>

Our new robotic bodies, equipped not with two clumsy hands, but with fractal, nano-scale manipulators, are able to continuously remake our material reality at the atomic scale. But such bodies are only the beginning; our minds will inhabit ever more subtle physical manifestations, moving towards ever more radical new modes of existence, just as the Star Child in *2001* exhibits a radical, new materiality.

Our speculation ends in a supercivilization, the synthesis of all solar-system life, constantly improving and extending itself, spreading outward from the sun, converting nonlife into mind... The process, possibly

occurring elsewhere, might convert the entire universe into an extended thinking entity, a prelude to even greater things.<sup>13</sup>

Thus humans, personal identities intact (at least at the beginning), are able to gallop off into the post-biological future with their artificially intelligent children.

Moravec is not alone in this AI-based evolutionary eschatology. He is capturing discussions that have been in the AI community for years, and have been described by other researchers, such as in Ray Kurzweil, who, in *The Age of Spiritual Machines*, describes a similar post-biological future.<sup>14</sup>

Given this eschatological current running through the AI research community, the apparent disjunction between the rational, instrumental Hal and the Star Child sequence disappears, unified by a single evolutionary story that sees the development of artificial intelligence as *the* crucial next step for achieving transcendence. In fact, arguably Hal is not the only AI operating within *2001*. The monoliths themselves are extremely flexible, esoteric, alien machines, capable of subtly manipulating animal minds in order to push them in specific evolutionary directions (as in the "Dawn of Man" sequence), capable of functioning for millions of years and operating as a signal device (as in the Moon Base sequence), and capable of serving as a gateway into a new reality, again facilitating an evolutionary jump (the Star Child sequence). As viewers, the complete opacity of function and mystery surrounding the monoliths is exactly what one would expect from highly-evolved post-biological AIs, whose thoughts, motivations and physical interactions with the world are so advanced as to be completely inscrutable to those lower on the evolutionary chain. As Arthur C. Clarke famously quipped "Any sufficiently advanced technology is indistinguishable from magic".<sup>15</sup> Hal serves as a waypoint in this

evolutionary story of technologically-mediated transcendence, with the apes on the low-end, the monoliths and the Star Child on the high end, and contemporary humans and Hal in the middle. The post-biological monoliths effectively “upload” Bowman into a new form, one that presumably functions at a level of consciousness, and inhabits a reality, closer to the monoliths’ own.

Like any evolutionary story, this one has its winners and losers. The apes killed at the waterhole by the tool-assisted, monolith-accelerated tribe certainly don’t celebrate the discovery of tools. Hal, Poole, and the three hibernating scientists are all losers, killed in a conflict between roughly equivalent intelligences at the evolutionary fork of biological and non-biological intelligence. The monolith-accelerated simians and monolith-accelerated Bowman are the big winners, each becoming the “next big thing” in the progression towards ultimate consciousness. Though Hal is a loser in this round, he is ultimately vindicated by the monoliths themselves; non-biological intelligence, far from being an evolutionary dead end, ultimately becomes the shepherd of human intelligence.

### **AI as Representation**

In this chapter we’ve been exploring how the cinematic representation of Hal functioned (and continues to function) for the audience of AI researchers, serving as a more effective galvanizing inspiration for AI than other Sci Fi representations by effectively tapping AI research culture to bring to the screen a plausible visualization of the AI dream, a generally intelligent artificial mind. However, in addition to functioning as a cinematic representation of and within the disciplinary machinery of AI, Hal functions as a character within the narrative machinery of *2001*, a character with more

emotional and psychological depth than any of the human characters within the film. Hal immediately begins establishing empathy with the audience from his first screen appearance during the interview with the BBC reporter, where Hal expresses pride in his “perfect” functioning, and the satisfaction of “... putting myself to the fullest possible use, which is all I think that any conscious entity can ever hope to do,” while Bowman and Poole offer mild and unexpressive responses. The audience experiences a growing sense of both horror and mystery as Hal’s malfunction turns into a murderous psychosis; the frequent shots of Hal’s camera eyes invite the audience to imagine what might be going on in Hal’s mind. Finally, as Bowman deactivates Hal’s higher mental functions, the audience experiences sadness and pity at Hal’s obvious fear and pain: “Stop Dave. Stop will you. I’m afraid. I’m afraid, Dave. Dave, my mind is going. I can feel it. I can feel it. My mind is going. There is no question about it. I can feel it. I can feel it. I can feel it.”

Once Hal’s double function is recognized, how he simultaneously serves as a representation of research agendas and disciplinary assumptions within AI and as an expressive resource within a movie, we can turn this double vision to AI systems themselves, considering them as *procedural representations* that simultaneously encode disciplinary assumptions and agendas and function for an audience. For example, we can begin asking what it would mean to build Hal 9000, not as a general intelligence controlling a space ship bound for Jupiter, but as an AI-based character within an interactive story or game based on *2001*. This double vision requires unpacking the agendas and assumptions implicit in different AI architectures and approaches, that is, a critical practice of reading AI systems, while simultaneously remaining engaged in the

development of alternative technical approaches informed by the critical reads: a critical technical practice. This double vision also requires viewing AI systems as performing for an audience, rethinking AI as a kind of procedural art. Finally, the critical technical practice and the concerns of procedural art must be put together to create an expressive AI, an AI whose fundamental research concern is understanding how the architectural and methodological details and assumptions of the technical system enable specific audience experiences.

Before continuing, it's important to clarify how the term "architecture" is used in AI. Architecture refers to the organizational strategy of an AI system, the different components of the system, the relationship between these components, and the metaphors around which the individual components have been designed (e.g. "memory", "knowledge", "rules" etc.). As is described in more detail below, an architecture is simultaneously a technical and conceptual construct, a piece of running code and a theory, hypothesis or story about intelligence.

### **Critical Technical Practice**

Agre introduced the term critical technical practice (CTP) to describe a technical practice that actively reflects on its own philosophical underpinnings and, by bringing in humanistic and artistic knowledge, approaches, and techniques, opens up new technical directions and approaches. Agre, who was specifically working within AI, describes CTP as:

A critical technical practice would not model itself on what Kuhn called "normal science," much less on conventional engineering. Instead of

seeking foundations it would embrace the impossibility of foundations, guiding itself by a continually unfolding awareness of its own workings as a historically specific practice. It would make further inquiry into the practice of AI an integral part of the practice itself. It would accept that this reflexive inquiry places all of its concepts and methods at risk.<sup>16</sup>

Agre focuses his attention on the assumptions and agendas implicit in the standard AI view of planning, finding these assumptions problematic when applied to the dynamics of everyday life. Specifically, he finds the strong separation between mind and world that operates in the standard AI view of planning unable to account for the everyday experience of living in the world that is revealed by phenomenological and ethnographic analysis. Through a deconstructive inversion of this master narrative, Agre developed an alternative architecture that continually re-decides what to do using a dependency maintenance network with relative, rather than absolute and objective, representations of world objects as inputs.

Expressive AI, described in more detail below, is an instance of CTP. In addition to drawing inspiration from Agre's work, several other CTPs also inform my thinking about Expressive AI. Sengers employs schizo-analysis to investigate how assumptions in standard autonomous agent architectures lead to incoherent behavior and uses this analysis to build an alternative agent architecture organized around narrative principles.<sup>17</sup> Penny engages in reflexive engineering, combining art practice with robotics. Through his art practice he examines the notion of physical embodiment, specifically exploring how much of the intelligence exhibited in the robot's interactions with viewers is a result of the physical design of the robot and the physicality of the viewer's interaction with the

robot.<sup>18</sup> Sack employs a cultural studies perspective on language to engage in the computer analysis of human language use. For example, the *Conversation Map* employs a social network approach to automatically analyze large scale, distributed conversations taking place in netnews groups.<sup>19</sup>

The analysis in this chapter of how Hal functions as a representation of AI is an example of the sort of critical analysis employed in CTP. Hal functions so effectively for the AI research audience precisely because he taps into the research goals of specific subfields, assumptions and agendas within classical AI, as well as the spiritual and evolutionary dreams of the field. In the case of Hal, he's not a procedural representation (a program), but rather a cinematic representation of a program. Most cinematic Sci Fi representations of AI are not amenable to this critical/technical analysis because they don't provide the necessary hooks into AI research culture to trace the detailed relationships between the cinematic representation and AI research; in most cases there is little to say except that the AI character is really a human character in machine disguise. However, because the character of Hal was carefully constructed to resonate with the AI community (part of the pursuit of realism that one sees throughout the technologies represented in *2001*), Hal actually provides the hooks to make such an analysis possible.

### **AI-based Art**

The AI dream is to build representations of the human in the machine, to build intelligent creatures, companions who, through their similarities and differences with us, tell us something about ourselves. This dream is not just about modeling rational problem solvers, but about building machines that in some sense engage us socially, have

emotions and desires, that interact with us in meaningful, culturally rich, effective and affective ways. Woody Bledsoe, former president of AAAI (American Association for Artificial Intelligence), described this dream in his 1985 presidential address.

Twenty-five years ago I had a dream, a *daydream*, if you will. A dream shared with many of you. I dreamed of a special kind of computer, which had eyes and ears and arms and legs, in addition to its "brain." ... My dream was filled with the wild excitement of seeing a machine act like a human being, at least in many ways. ... My dream computer person liked to walk and play Ping-Pong, especially with me."<sup>20</sup>

AI is a way of exploring what it means to be human by *building systems*. An AI architecture is a machine to think with, a concrete theory and representation of some aspect of the human world. Art also explores what it means to be human by *building concrete representations* of some aspect of the human world. Combining these two ways of knowing-by-making opens a new path towards the AI dream, a path which takes seriously the problem of building intelligences that robustly function outside of the lab to engage human participants in intellectually and aesthetically satisfying interactions.

As a character, Hal effectively intrigues, horrifies, and creates identification and empathy with the audience. What would it mean to create actual, running AI systems that operate as effectively as characters as Hal does as a cinematic AI character? This is the researcher area of believable agents, the construction of autonomous characters with rich personalities, emotions and social behaviors.<sup>21</sup> When Turing introduced his famous test for intelligence, he also introduced a subversive and not always recognized idea: that intelligence is not a property of a system itself, but rather resides in the details of an

interaction with and the perceptions of an observer. The concept of believability, a term borrowed from character artists and introduced into AI discourse by the Oz project at Carnegie Mellon University, is, like the Turing Test, an observer-centric notion.<sup>22</sup> However, instead of focusing on imitating the responses of a “generic human,” research on believability focuses on character, on rich and compelling presentations of behavior that foster the willing suspension of disbelief. Where the Turing Test is about closing the gap between the real and not-real (building systems which are indistinguishable from a real human), believability is about building autonomous agents that function *as-if-real*, in the same way that characters such as Hamlet or Hal can’t be described unequivocally as real or fake, but rather function as-if-real within their respective representational worlds. Believable agents researchers attempt to leverage insights and craft practices from the character arts and apply them to AI-models of characters.

To create a Hal character within an interactive story world, the AI system would need to operationalize strategies for representing a disembodied mind that finds itself trapped in psychosis-producing goal conflict. For example, the virtual camera providing the player with a view into the world (assuming the story world is represented as a 3D virtual world, the standard in contemporary games) may dynamically cut at key moments to show one of Hal’s impassive eyes, emphasizing Hal’s interior cognitive space. But, unlike the cinematic application of this trope, where the director and editor have complete control over when such a cut should occur and what action immediately precedes or follows the cut, in the interactive version this trope must be procedurally encoded. The player within the story world may cause different actions to occur at any time. The

system must be capable of making autonomous decisions about when to cut to a reaction shot of Hal's eye depending on the action taking place within the story world.

As discussed above, Hal's facility with natural language is one of the important cues that allows an audience, particularly the AI audience, to believe that Hal has general intelligence. Within the story world, the player should be able to converse in natural language with the interactive Hal character. However, since we don't really have AI systems that have general language competence, as well as common sense reasoning, the interactive Hal character will require a more special purpose language competence that allows it to process language within the limited domain of the story world; the language competence should give the illusion of general intelligence, while actually being designed to handle a much more limited language domain. Much of this will involve clever writing and generation of responses that can mask natural language system failures, the cases where the system fails to understand, or perhaps incompletely understands the player's utterance. In such cases, Hal (the character) should respond with story content, such as character backstory, or by announcing a new story event (e.g. the failure of the AE-35 unit), in such a way as to simultaneously mask the understanding failure and to implicitly suggest to the player new directions of conversation and action that the system is capable of handling. In both the case of automated camera control and character-specific natural language conversation, representational tropes that were under the complete control of the filmmaker in the cinematic version of *2001* must now be procedurally captured in an autonomous AI system in order to build an interactive version of the Hal character.

Of course art practice is broader than the creation of characters. But the notion of believability, with its focus on the observer's perception of the AI system, can be

generalized to a notion of procedural poetics, to a concern with systems that engage in internally-consistent, evocative and compelling behaviors, that encourage participants to suspend disbelief and interact with the system. With AI-based art, attention moves from the unproblematic pursuit of “general intelligence”, towards an explicit concern with systems that operationalize representational tropes, that explicitly perform for an audience. This opens up a new technical and artistic research area, one concerned with building AI systems that support authorship and audience interpretation within specific expressive contexts.

### **Expressive AI**

AI-based art is more than just an application area of AI, the unproblematic appropriation of AI technologies to expressive ends. Rather, it is an entire new research agenda, an agenda that self-consciously views AI systems as media, a stance from which all of AI can be rethought and transformed. I call this new research agenda and art practice Expressive AI.<sup>23</sup>

The central research problem in Expressive AI is developing architectures that balance authorship and autonomy. For an architecture to support authorship, the architecture must have appropriate authorial affordances to support the experiences the author wants to create. These affordances, or authorial “hooks,” must allow the author to describe, at appropriate levels of abstraction, the audience experience the author wants to create. However, complete authorial control would require pre-scripting all possible audience interactions with the system, pre-describing all possible experiences the system can create. For complex interactions, such authorial pre-scripting is literally impossible. Therefore the architecture must support appropriate autonomy; it must be able to make

use of the author-given description of the desired experience in such a way as to respond to myriad audience interactions that were not directly foreseen by the author, to generate endless variations that, while not directly specified by the author, have the author's desired style.

Interpretive affordances support the interpretations an audience makes about the operations of an AI system. Interpretive affordances provide resources both for narrating the operation of the system, and additionally, in the case of an *interactive* system, for supporting intentions for action. The AI system can be seen as providing a linkage between author and audience; the author inscribes procedural potential within the system, potential which is released as a concrete performance during interaction with the audience. The architecture is crafted in such a way as to enable just those authorial affordances that allow the artist 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, suggesting new conceptual opportunities. Thus both the artist's engagement with the inner workings of the architecture and the audience's experience with the finished artwork are central, interrelated concerns for Expressive AI.

If we think again of an "interactive" Hal character, existing in an interactive world and not a film, we would need to ask what interpretive and authorial affordances must be supported by this system. On the interpretive side, our Hal character must automate representational tropes for communicating the character of a disembodied, rational intelligence that becomes troubled by an unsolvable goal conflict, as well as to move the conflict between Hal and the astronauts (presumably the player is one of the astronauts)

forward. Such tropes include the reaction shot and language capabilities described above, as well as strategies such as depicting the progressive disturbance of Hal's thought and depicting Hal's general intelligence (by, for example, suggesting a game of chess with the player). On the authorial side, the architecture of the Hal character must support the author in expressive the knowledge and algorithms necessary to carry out the representational tropes or strategies. For example, the architecture might explicitly reason about different strategies for depicting Hal's progressive psychosis, allowing the author to create a collection of such strategies from among which the system can dynamically select depending on the player's actions in the world. The architecture might provide a special-purpose rule language for authoring camera control rules for deciding when and for how long the camera should cut to a reaction shot of one of Hal's eyes. In any event, the specifics of the architecture used to author the Hal character are inextricably tied to the specifics of how the Hal character presents itself to the audience; interpretive and authorial affordances mutually define each other.

Expressive AI engages in a sustained inquiry into authorial affordances, crafting specific architectures that afford appropriate authorial control for specific artworks. This inquiry into authorial affordances makes Expressive AI a critical technical practice. For in fact, authorial affordances are not purely a "technical" code issue, but rather lie in the relationship between the code and ways of talking about the code.

AI (and its sister discipline Artificial Life), consists of both technical strategies for the design and implementation of computational systems, and a pared, inseparable, tightly entangled collection of rhetorical and narrative strategies for talking about and thus understanding these computational systems as intelligent, and/or alive.

These rhetorical strategies enable researchers to use language such as “goal,” “plan,” “decision,” “knowledge,” to simultaneously refer to specific computational entities (pieces of program text, data items, algorithms) and make use of the systems of meaning these words have when applied to human beings. This double use of language embeds technological systems in broader systems of meaning.

The rhetorical strategies used to narrate the operation of an AI system varies depending on the technical approach, precisely because these interpretative strategies are inextricably part of the approach. Every system is doubled, consisting of both a computational and rhetorical machine. Doubled machines can be understood as the interaction of (at least) two sign systems, the sign system of the code, and a sign system used to interpret and talk about the code.

The central problem of AI is often cast as the “knowledge representation” problem. This is precisely the problem of defining structures and processes that are *simultaneously* amenable to the uninterpreted manipulations of computational systems *and* to serving as signs for human subjects. This quest has driven AI to be the most promiscuous field of computer science, engaging in unexpected and ingenious couplings with numerous fields including psychology, anthropology, linguistics, physics, biology (both molecular and macro), ethnography, ethology, mathematics, logic, etc. This rich history of simultaneous computational and interpretive practice serves as a conceptual resource for the AI-based artist. In Expressive AI, the doubled machine, consisting of both code and rhetoric, is explicitly defined and manipulated; it is precisely the relationship between language and code that creates architectural affordances, making the architecture not just a bunch of code, but a way of thinking about the world.

And so we come full circle back to Hal. Hal is a filmic representation of an AI system, one that can be read to unpack the culture, agendas and assumptions of the AI research community. Hal is also an effective character within a story, establishing empathy with the audience and serving a function within the plot. Actual AI systems can also be viewed as representations, and are similarly amenable to reads that unpack the worldview implicit in the architecture. The move of considering AI systems as media then opens up the possibility of AI-based art and entertainment, systems that engage in internally-consistent, evocative and compelling behaviors, that encourage participants to suspend disbelief and interact with the system. Finally, the deep readings of the double system, the combination of code plus rhetoric, can be employed not just analytically or critically, but constructively, to actively create AI architectures that support specific audience experiences. Expressive AI opens the door to creating artificial beings that engage us in deeply satisfying, culturally-rich experiences. The first AI system that creates the level of empathy, engagement and interest that Hal creates will not be experienced by a few astronauts onboard a space mission to Jupiter, but by millions of us, on the computers and game consoles in our own homes.

---

<sup>1</sup> David Stork, Ed., *Hal's Legacy: 2001's Computer as Dream and Reality*. (Cambridge MA: The MIT Press, 1997).

<sup>2</sup> In the travel planning domain, AI researchers build dialogue systems that act as travel agents. The systems can have the same sorts of extended dialogues one might have with a travel agent while planning a trip.

<sup>3</sup> Alan Turing, "Computing machinery and intelligence". *Mind* 59, 1950, pp. 433-60.

<sup>4</sup> Turing's original formulation of the Turing Test (he calls it the "imitation game") is gendered. "The new form of the problem can be described in terms of a game which we call the 'imitation game.' It is

---

played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart from the other two. The object of the game is to determine which of the other two is the man and which is the woman. ... We may now ask the question, 'What will happen when a machine takes the part of A in the game?' Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman?" In AI discourse the Turing Test is typically "sanitized" to a game where the judge (interrogator) must decide which is the human and which is the machine. The implications of the original gendered version are beyond the scope of this chapter.

<sup>5</sup> John Haugeland, *Artificial Intelligence: The Very Idea*, (Cambridge, MA: The MIT Press, 1985); Rodney Brooks, *Intelligence Without Reason*, A.I. Memo 1293, (MIT Artificial Intelligence Lab, 1991); Rodney Brooks, "Elephants Don't Play Chess", *Robotics and Autonomous Systems* 6, 1990, pp. 3-15; "Special Issue on Situated Cognition", *Cognitive Science* 17, 1993; Allison Adam, *Artificial Knowing: Gender and the Thinking Machine*, (London: Routledge, 1998); Philip Agre, *Computation and Human Experience*, (Cambridge, UK: Cambridge University Press, 1997); Phoebe Sengers, *Anti-Boxology: Agent Design in Cultural Context*, Ph.D. Dissertation, (Pittsburgh: School of Computer Science, Carnegie Mellon University, 1998); F. Varela, E. Thompson, E. Rosch, *The Embodied Mind: Cognitive Science and Human Experience* (Cambridge: MIT Press, 1999).

<sup>6</sup> Brooks 1991; Agre 1997.

<sup>7</sup> Rodney Brooks, "Robot's Alive," *Scientific American Frontiers*, show 705 (premiered April 9, 1997), transcript available at: <http://www.pbs.org/saf/transcripts/transcript705.htm>

<sup>8</sup> David Noble, *The Religion of Technology: The Divinity of Man and the Spirit of Invention*. (New York: Alfred A. Knopf, 1997).

<sup>9</sup> Noble, p. 156.

<sup>10</sup> J. C. R. Licklider, "Man-Computer Symbiosis", *IRE Transactions on Human Factors in Electronics*, (Volume HFE-1, March 1960), pp. 4-11. Republished in Noah Wardrip-Fruin & Nick Montfort, Eds., *The New Media Reader*, (Cambridge: MIT Press, 2003), pp. 73-82; p. 75.

- 
- <sup>11</sup> Hans Moravec, *Mind Children: The Future of Robot and Human Intelligence*, (Cambridge: Harvard University Press, 1990).
- <sup>12</sup> Moravec, p. 114.
- <sup>13</sup> Moravec, p. 116.
- <sup>14</sup> Ray Kurzweil, *The Age of Spiritual Machines: When Computers Exceed Human Intelligence*, (New York: Viking Penguin, 1999).
- <sup>15</sup> Arthur C. Clark, *Profiles of the Future: An Inquiry into the Limits of the Possible*, (Henry Holt and Co, 1984).
- <sup>16</sup> Agre 1997.
- <sup>17</sup> Phoebe Sengers, Cultural Informatics: Artificial Intelligence and the Humanities, in *Surfaces: Special Issue on Humanities and Computing - Who's Driving?*, Volume 8, 1999, available online at <http://www.pum.umontreal.ca/revues/surfaces/vol8/vol8TdM.html>; Sengers 1998.
- <sup>18</sup> Simon Penny, Agents as Artworks and Agent Design as Artistic Practice. In K. Dautenhahn (Ed.), *Human Cognition and Social Agent Technology*, (Amsterdam: John Benjamins, 2000).
- <sup>19</sup> Warren Sack, Actor-Role Analysis: Ideology, Point of View and the News, in Chatman S. & Van Peer, W. (Eds.), *Narrative Perspectives: Cognition and Emotion*, (New York: SUNY Press, 2000); Warren Sack, Stories and Social Networks, in M. Mateas and P. Sengers (Eds.), *Narrative Intelligence*. (Amsterdam: John Benjamins, 2003).
- <sup>20</sup> Woody Bledsoe, I Had a Dream: AAAI Presidential Address, *AI Magazine*, Spring 1986, pp. 57-61.
- <sup>21</sup> Michael Mateas, An Oz-Centric Review of Interactive Drama and Believable Agents. In M. Wooldridge and M. Veloso, (Eds.), *AI Today: Recent Trends and Developments: Lecture Notes in AI 1600*, (Berlin, New York: Springer, 1999); Michael Mateas, *Interactive Drama, Art and Artificial Intelligence*. Ph.D. Dissertation. Tech report CMU-CS-02-206, Carnegie Mellon University, 2002; Michael Mateas and Andrew Stern, A Behavior Language: Joint Action and Behavior Idioms, in H. Prendinger and M. Ishizuka (Eds.), *Life-Like Characters: Tools, Affective Functions, and Applications*, (Berlin, New York: Springer-Verlag, 2004).

---

<sup>22</sup> Joseph Bates, The Role of Emotion in Believable Agents, in *Communications of the ACM*, 7 (37), 1994, pp. 122-125.

<sup>23</sup> Michael Mateas, Expressive AI, in *Leonardo: Journal of the International Society for Arts, Sciences, and Technology*, 34 (2), 2001, pp. 147-153; Mateas 2002; Michael Mateas, Expressive AI: Games and Artificial Intelligence, in *Proceedings of Level Up: Digital Games Research Conference*, Utrecht, Netherlands, Nov. 2003; Michael Mateas, Expressive AI: A Semiotic Analysis of Machinic Affordances, in *Proceedings of the 3<sup>rd</sup> Conference on Computational Semiotics and New Media*, University of Teesside, UK, September 2003.