10th International Conference on Computer Graphics and Artificial Intelligence, Athens, Greece, May 30-31, 2007, 123-134.
Keynote presentation, 3IA 2007

Avatar Intelligence Infusion
– Key Noteworthy Issues

Kristinn R. Thórisson

Associate Professor, Department of Computer Science
Co-Director, Center for Analysis & Design of Intelligent Agents
Reykjavik University
Kringlunni 1, 2nd fl., 103 Reykjavik, Iceland
thorisson@ru.is

Two main reasons exist for endowing graphical agents with artificial intelligence: To enable them to act as fully autonomous, independent, cognitive agents and to help humans control their behavior more easily. Both are a challenge to researchers, programmers and designers in computer graphics and artificial intelligence. In this paper I discuss some of these challenges and propose steps for addressing them. I claim that a prerequisite for building autonomous and human-controlled avatars is a solid implementation of a perception-action loop and argue for the need to pay equal attention to perception as to action. I discuss the concept of *cognitive presence*, which can be used to evaluate both human-controlled avatars and autonomous agents, and highlight the nature of these concepts and explain their place in the development of computer games through examples from natural human interaction. I present arguments in support of the claim that significant architectural elements can be shared between avatars and fully autonomous agents and that in both cases it will be difficult to achieve high levels of cognitive presence without proper architecture.

Introduction

The field of game design has achieved significant results in the past two decades, particularly in the areas of graphical realism and rendering speed. Increasingly, development houses are looking towards greater realism of the players in the game, both the graphical representation of human players and the behavior of fully autonomous agents – so-called non-player characters (NPCs). At first glance these two tasks may seem to be completely different; the first concerns human interface and graphical techniques, the second concerns artificial intelligence. However, on closer inspection the two goals have much more in common than one might think. To create graphical characters – avatars – with intuitive controls for a human takes a lot more than clever mapping of keyboard commands to the body of the avatar; its body has dozens of degrees of freedom that need to be controlled intuitively in dynamic conditions. It turns out that this is very difficult to do seamlessly without complex transformation rules between the input and the output, calling for sophisticated technologies that interpret simple input, like a joystick's movements, into multi-dimensional movements of a body in a real environment. Some of the challenge concerns the necessity to limit the input of game players to relatively simple input devices and some of it has to do with the fact that most game playing experience is achieved through the exclusion of certain realism, e.g. the need to control the legs of a walking character, in return for increased focus on more cerebral aspects of the game. This is where fully autonomous technologies and collaborative control meet: The complexity of creating a system where human and machine manage complex movements together. The trick is to make the outcome believable, enjoyable and in support of the game's goals.

In this paper I explore the connection between avatars and intelligence from the perspective of *cognitive presence* – the subjective second- and third-person experience of intelligence, whether mediated, real or synthetic. First I describe the concept itself shortly. Then I discuss key topics in intelligence and interaction: feedback and its relation to autonomous and semi-autonomous humanoid agents. I relate the issues of intelligence architecture and cognitive architectural validity to cognitive presence and draw conclusions about their combined importance in the design and development of autonomous agents and human-controlled avatars.

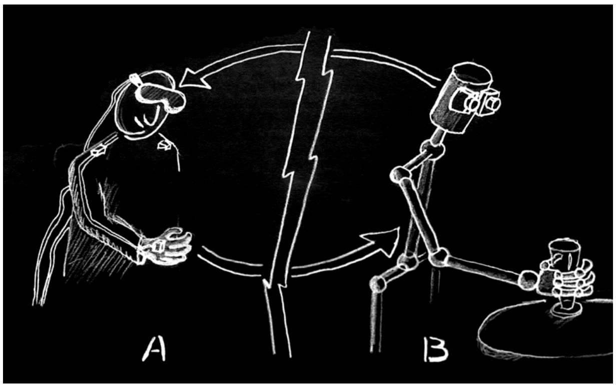


Figure 1. *Telerobotic system: Teleoperator (A) with telerobot (B). Control signals (lower arrow) from teleoperator (A) are carried to the robot's body (B); sensory signals – vision, hearing, touch – are carried back from the robot to the operator's goggles, headphones and gloves (upper arrow). The tightness of this perception-action loop (combined arrows) determines in great part the experience of presence: The more direct the coupling and the lower the transmission delay, the stronger the sense of telepresence experienced by the operator.*

Cognitive Presence

How can you be sure that the letters streaming back through your instant messaging terminal are actually being typed by a person at the other end of the line? If you sense that a thinking mind is behind the words you are reading you experience *cognitive presence*. The higher the level of cognitive presence experienced, the more likely it is that the typist is in fact a human.

Cognitive presence is an extension of the concept of telepresence, familiar to those working in the field of telerobotics (cf. Goldberg 2000, Held & Durlach 1992, Sheridan 1992). In this field presence refers to the level of perceptual immersion and emotional involvement that an operator of a remote robot feels as he operates the device. Typically cameras provide visual information and sometimes force feedback is used to provide

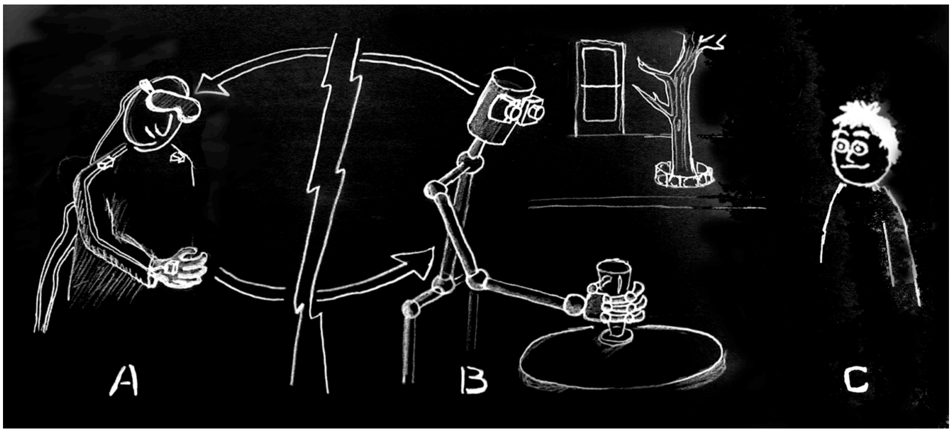


Figure 2. *An onlooker (C) in a telerobotic setup may experience the telerobot as expressing* cognitive presence *if the robot's actions contain features that the onlooker sees as being caused by a thinking being with cognitive processes. The level of cognitive presence felt is a function of, among other things, the coherence of the observed entity's actions, which in turn depend directly on the coherence of its mental architecture (see text).*

haptic cues (Figure 1). A sense of presence means that the operator feels as if she is present at the location of the robot, not the location where her own body resides (which, by the definition of a telerobotic system, must be at a location remote from the robot).

It is ultimately the ability of the system to evoke natural, entrenched responses to the various mediated circumstances that determines the level of presence induced by a telerobotic system: The higher the fidelity of the sensory feedback and control coupling between the operator and the robot, the deeper the sense of presence felt. Standard tricks for achieving a strong sense of presence, such as stereoscopic goggles, low latency of signals and other such gadgets, are not of interest here however; what we are interested in is a third element – *the observer* (C in Figure 2). Cognitive presence thus extends the notion of telepresence by introducing an observer – a third party – into the scene. The observer is looking at the telerobot and trying to understand what it's doing. His sense of understanding, his sense of coherence in the robot's actions, and the robot's similarity to himself, all contribute to an experience that in the onlooker results in an evaluation of the robot's *cognitive presence*. Cognitive presence is defined as *an observer's sense of thought being present in an animate entity* – the feeling that "somebody is home".

One could also define cognitive presence as the *sensed evidence for mental processes causing the observed behaviors*. As shorthand we will say that an entity "has presence", and "is capable of expressing presence", if it has the ability to evoke a sense of cognitive presence in an onlooker. We will return to this concept below.

An intuitive way to look at the phenomenon of cognitive presence is in the context of human intelligence: As this is the kind of intelligence we are most familiar with (being its embodiment, as it were) we naturally use this to judge other intelligences by. Cognitive presence will thus be strongest if the behaving entity is similar in form and function to ourselves – to a human.

The Perception-Action Loop

With today's computing power it seems we should be able to give computer-generated characters a sense of lifelikeness and at least some noticeable amount of cognitive presence. NPCs in computer games present a practical target. What complicates things, however, is that any system that is to display a sense of self, by being aware to some extent of the world around it, needs to be imbued with some sort of perception: The sense of its own actions, should they fail to get executed; a sense of the environment, should it prevent its actions. Actions need not only take into consideration the goals of the character, they need to take into consideration unexpected events of importance to the character. To start, this requires a perception-action loop (Figure 3). Innocuous as it may seem, this initial step of implementing a solid loop has made life hard for even the most accomplished developers. The reasons for this are not well understood, but some answers are beginning to emerge. Let me try to explicate some of them.

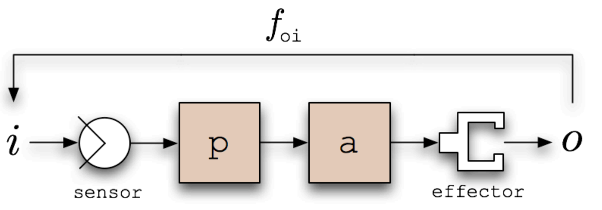


Figure 3. *The outer feedback loop in any self-monitoring system is the one that has been the hardest to model in computer graphics systems. This loop is the same as that shown in Figure 1. (Key: p: perception process, a: action and planning process, i: external input, o: generated output and environmental effects, f_{oi}: outermost feedback loop in any intelligent system, from environmental effect to sensation via external sensory mechanism. Notice that the diagram abstracts the cognitive architecture to input and output functionality only, and is therefore a significant simplification).*

It turns out that the perception-action feedback loop is key in separating dead objects from live ones. The failure of a falling rock to make *any* preparations in advance of the fact that it is about to hit the ground at 200 kilometers per hour gives it its very rockness – its gross anatomical behavior as it splinters on the ground is subject to laws that have been known for centuries and can be automatically computed in a feed-forward manner. Granted, collision detection has to be dealt with in the impact and as the splinters scatter and bounce on the ground, but these are also the result of repeated feed-forward processes that can be quantified with a few parameters – in short, the math is understood. Not so for interactive cognitive systems.

As we further study the feedback loop in the context of humans it dawns on us that it is not only necessary to take its simplest version into account – there exist in fact many loops, at varying levels of complexity. The range of time that they span covers several orders of magnitude, from reflexes to planning. Knowing where to start – and where to stop – when implementing these is typically a major challenge, even in the simplest game environments.

Figure 4 shows some of the feedback loops that an intelligent, learning being needs. The most commonly discussed feedback loop in humans, the reflex, is actually not interesting enough to enter into this picture.[1] Here we are more interested in the voluntary feedback loops: The first is the quickest, so-called *simple reaction time* – the highest speed at which we can make a voluntary movement, e.g. pressing a button, at the cue of an external stimulus, e.g. a bell (cf. Luce 1986). The second is on average at the task level – roughly the speed that we e.g. can make corrections during everyday tasks based on simple recognition of cues (for example that a door opens out instead of in, or the speed with which we notice and correct an incorrectly pronounced word). The third is the kind where we need to make deeper inferences about causes and effects, e.g. inferring that the washing machine doesn't turn on because its power cord has been shaken loose, or the particular strategy we intend to use to persuade someone to buy the house we want to sell, to take two examples from everyday life.

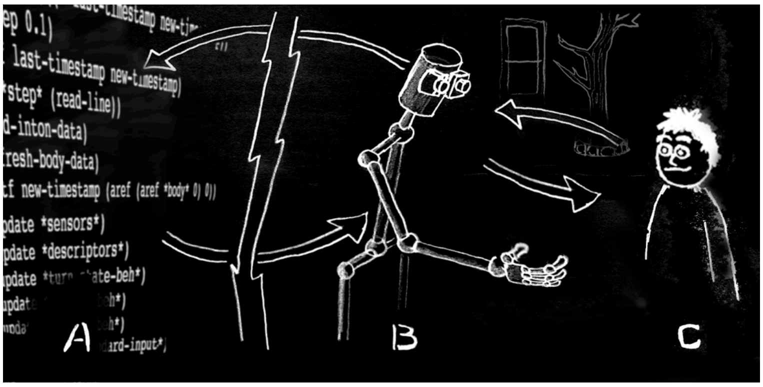
Cognitive Presence Mediation

When observing natural biological systems, cognitive presence is evoked by how closely the observed dynamic features, or behaviors, resemble those observed in other systems known to possess cognition, especially human forms. The same rule holds when observing unknown systems, whether natural or artificial. Thus, a number of indicators combine to make the overall impression of cognitive presence. The strength of the presence experienced is a function of the underlying thought processes of the observed system, but is also limited by the ability of the system's underlying processes to express their existence via some recognizable medium such as a familiar body shape.

We can differentiate between *statically mediated* and *dynamically mediated* cognitive presence. Into the former class fall such phenomena as an abandoned house, a carefully crafted tool, and a painting: All convey some sort of thinking, of cognitive effort expended – and its presence in the creation of the artifact can be felt at a later time. Dynamically mediated cognitive presence can be classified

[1] Neural signals in the reflex arc never reach the brain, they only travel from the point of stimulation to the spine and back to the muscles. They are therefore as close to a hard-wired feature as we get in the body, and can appropriately be implemented with simple IF-THEN rules.

into *non-interactive* and *interactive*. Into the former class fall phenomena such as a video clip of a being moving about or doing something, recorded speech, a musical performance and even the sounds of goal-oriented tool usage. *Interactive cognitive presence* is a sense of cognitive presence evoked through interaction with an intelligence, e.g. through words exchanged via instant messaging. *Embodied cognitive presence* is the sense of cognitive presence evoked by the behavior of an actual physical embodiment of a behaving intelligence. Interactive cognitive presence does not have to imply embodied cognitive presence, or vice versa: As in the case of a letter received in response to our own letter, the interaction can be significantly displaced in time and happen via different media. And finally, *Embodied interactive cognitive presence* is created by interaction with the physical manifestation of a behaving intelligence. This is the cognitive presence we experience when interacting with others face-to-face.



```
ep 0.1)
  last-timestamp new-timestamp)
*step* (read-line))
d-inton-data)
resh-body-data)
f new-timestamp (aref (aref *body* 0) 0))
pdate *sensors*)
pdate *descriptors*)
update *turn-state-beh*)
updat          h*)
upda           h*)
               ard-input*)
```

Figure 5. *Experience of cognitive presence may be produced in a human observer (C) by an autonomous system (A). A strong sense of presence is difficult to achieve this way, especially if the human can interact with the system's embodiment. Depending on circumstances, the cognitive presence may vary in strength, from a powerful cognitive presence to a weak one, as many features may affect it such as the context of the interaction, duration, goal and topic of discussion. However, the major determinant is the system's architecture.*

When thinking about these different classes of cognitive presence it is useful to imagine the boundaries – the cases where cognitive presence is either super-enhanced or significantly subdued. The former case includes interaction with awake humans: Their gaze, muscle movements, facial expressions, etc., are perfectly life-like and capable of conveying the strongest form of cognitive presence. If we watch animated characters in a movie, clever voices, music and other tricks can help create cognitive presence, but if we wish to convey a strong sense of presence to the viewers, a real human will always be a better choice than an automatically controlled graphical character, a mechanical puppet, or even a hand-controlled Muppet. On the other end of this spectrum is the case where cognitive presence disappears, even in cases where intelligence is present. What might that look like?

Let's go through the various types of presence mediation. *Static*: If aliens live in buildings that look like natural rock formations we would be inclined not to perceive any cognitive presence when examining them – whether on location or in photographs – no matter how smart they may have been (not even if we knew that the rocks were hand-made by aliens). Dynamic, non-interactive: If the aliens look more or less like a waterfall as they move about it will be difficult to distinguish them from the more familiar waterfalls on Earth. Interactive: We can interact with the aliens but they look like waterfalls and communicate by carefully modulating the water splashes, drop size and surface waves. In each of these cases we fail to experience any sense of cognitive presence. Notice, however, that this is not caused by lack of alien intelligence but rather our failure to recognize its form of expression.

Notice also that we are most unlikely to perceive cognitive presence in the static case and most likely in the interactive case. The reason is the increase in degrees of freedom: The interactive case provides us with a range of methods to do little "experiments" on the intelligence, probing how it responds to various stimuli. Through such interaction, discontinuities in the intelligence makeup, or of its expressive form, will quickly become apparent. Interaction through multiple modes greatly increases the number of degrees of freedom in the system and explodes the possible range of variations during such "experimentation". This is why *multimodal* *interaction* is among the most difficult forms of embodiment of artificial intelligence (Figure 5) – it calls for a highly coherent architecture that can coordinate a large amount of heterogeneous processes (Thórisson 2007).

For the remainder of this paper we will focus on embodied, interactive intelligence. In particular, I will discuss two things that I believe define human-like cognitive presence to a larger extent than most other: *Natural multimodal communication*, *turntaking* and *cognitive architecture*.

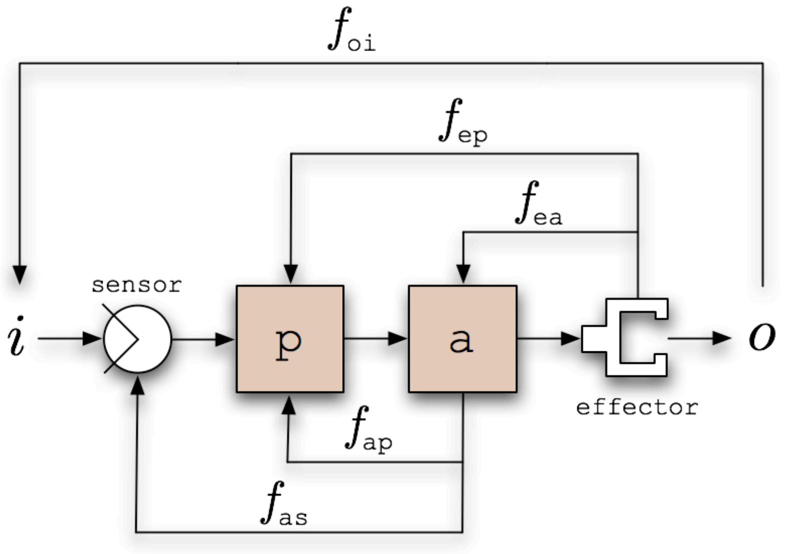


Figure 4. *A number of feedback loops enter into the operation of all cognition, in addition to the real-world feedback (f_{oi}). A feedback loop helps the action controller steer the end-effector (f_{ea}); another monitors the relationship between the actions performed and other perceptions (f_{ep}); a third one monitors the operation of the action controller (f_{ap}); yet another one helps the action controller steer the sensors (f_{as}). Depending on the particular implementation of the full architecture, removal of any one of these loops may change the cognitive presence of a character.*

Intelligence

The goal of NPCs and avatars in computer games is to evoke a sense of believability. Cognitive presence is an important part of this believability – an autonomous game character with little intelligence will sooner or later stick out like a sore thumb as one or more of its decisions, behaviors, memory failures, or other facilities with direct correspondences in the real world fail to match our expectations. Similarly, an avatar whose behavior requires its human user to control everything manually, from eye gaze to its walking motion to finger movements, will be drowned in information and control panels. Such an avatar would in fact not be able convey a sense of cognitive presence but rather would exude chaos and confusion. Therefore, some of the technologies we infuse into autonomous characters can alleviate the burden of control through automation while leaving certain parts under the control of the owner (Vilhjálmsson 2004, Cassell & Vilhjálmsson 1999).

Natural Multimodal Communication

Speech is sequential and one-dimensional. This anchors the amount of information we can convey via speech tightly to a realtime clock with an upper limit. Nature has provided various methods that lessens this limitation, primarily among them being our ability to use multiple modes for communication – in fact multimodal communication appeared millions of years before written and even spoken language, and is thus much more natural than unimodal communication. While language has mainly to do with the words we choose and how we order them, and conversely, how we interpret such streams of words, natural face-to-face communication additionally involves a complicated coordination of multi-dimensional bodies. Combining modes can easily create new meaning or change the meaning of the words we are saying.[2] Science has split the modes apart in its analysis of communication, primarily because of the advent of written language. The challenge is now to put the modes back together: To understand, for example, how it is that a facial expression or a glance of the eyes combines with an utterance to modify or completely change its meaning.

The Turing Test was proposed as a way to identify intelligence through language use (Turing 1950) but has been criticized as a tool for this purpose (cf. Hayes & Ford 1995).[3]

[2] For example, if we mutter “he’s such a great person” and roll our eyes, many would understand the sentence to mean the exact opposite of what the words alone convey.

[3] The Turing Test consists of an instant messaging game where a person sits on one end and either a computer program or a human at the other. The observer’s task is to

While the test is supposed to be about content, what is being said, rather than the process of interaction – and Turing himself certainly intended for it to work that way – some researchers have imbued their programs with typing skills that mimic that of humans (making typos, deleting back a few characters to correct words, etc.), as a way to better fool the observer. As it turns out, such tricks help quite a lot, in fact, to fool human observers intent on differentiating programs from humans: Clearly it is not only *what* we say that matters, but also *how* we say it, as any politician could have told us. This is a strong indication that rather than being a measure of general intelligence, the Turing Test actually measures *interactive*, *non-embodied cognitive presence*.[4]

Language generation and understanding is, by any measure, a complex phenomenon that is difficult to automate. However, systems have been built that are good enough to fool human game players for long periods of time, using typed language via instant messaging. Interestingly, as the development of computer games is open to a lot of creative manipulation of story, actors and environment, some of the challenges associated with implanting language skills in our autonomous agents can be reduced through clever game design. A greater problem turns out to be a follow-on requirement for any language-capable being: to be able to use language to refer to past events, states of its own being, prior interactions in social situations, deductions that any five year old could make, as well as remembering and forgetting such things in a way that is human-like. To do so they must be able to sift out that which is important from that which is not. At that point we are already knee deep in cognitive architecture and artificial intelligence. But before we delve into that further I want to discuss yet another set of challenges.

Turntaking, Gaze & Attention

When people communicate they take turns speaking. Turntaking is a universal feature of human interaction (Goodwin 1981). A lot of subtle yet important things happen during turntaking: Glances are exchanged, attention is directed, tone of voice, intonation and words are produced and interpreted. The orchestration of such events is quite involved and impressive, yet we do it so effortlessly that the complexity is easy to overlook. It is this interplay of modes, information types and communication preferences – orchestration that has been exceedingly difficult to model computationally – that creates the perception of cognitive presence. As a case in point, let's take gaze – a small yet extremely important part of conveying cognitive presence. As everyone knows, few things show lack of attention more clearly than someone's failure to look back at us right after we have told them big news. Gaze is a fairly direct indication of visual attention, and even of auditory attention (Riesberg et al. 1981). Attention, in turn, is our "internal CPU" which we must carefully control at every point in time so as to conduct our life – if we don't we are at risk of being hit by a car or inadvertently insulting our boss by being absent-minded. Gaze is controlled by a number of processes that have to be tightly coordinated, that includes selections of points of fixations 2-5 times per second (Card et al. 1986), looking in the relevant direction at the right time, avoiding insulting people by staring, gazing upwards to show you don't want to be interrupted while recalling a word (Argyle & Cook 1976), etc. (And remember that these movements have to be coordinated with other body movements and everything else that's going on – by both participants in the conversation.) Movements with such high frequency updates need very small deviations from the expectations of an experienced observer to be detected. We are all such experienced observers.

Turns in dialogue exist because of limitations of cognition – we are incapable of listening and speaking[5] at the same time for very long periods

interact with the other entity and determine whether it is a human or a computer. If the observer mistakes a computer program for a human then that program must necessarily be intelligent, according to the test.

[4] Further, because of the test calls for a comparison to a human, the test is clearly a test of *human* non-embodied interactive cognitive presence. See Harnad (1991) for extensions to the Turing Test.

[5] This applies to all forms of communication, not just speech. For convenience, however, we will use "speech" and "listening" to stand for the creation and interpretation of any

of time, a few seconds at most.[6] As gaze is closely related to the mental operations we perform relating to visual and auditory information gathering, gaze has been shown to play an important role in turntaking and communication. For example, speakers tend to look more at listeners and certain repeated patterns of fixations and saccades are associated with turn exchanges (cf. Kleinke 1986, Duncan 1972). This is because gaze is associated with certain cognitive events that have to do with keeping track of the dialogue topic (e.g. looking at objects under discussion), avoidance of distraction (e.g. looking away), as well as being trained in social situations to play certain roles (e.g. looking at the face of someone we have just asked a question).

These are some of the reasons that gaze and turntaking are important for cognitive presence. The key insight is that our eyes are controlled by a number of (relatively complex) processes and thus the gaze patterns observed when people communicate are an emergent property of interaction between these processes. Somehow conflicting goals, at multiple levels of detail, are resolved and a coherent, believable coordination of all modes emerges. The reason that it's believable is key, however: It is believable because of our expectations of what it is like to be that behaving being – we understand other people to a sufficient extent to be able to interpret and evaluate their behavior. To the extent that we can, we experience a cognitive presence in others. The main reason we are able to do this is because we share something with them – we think alike. The reason we think alike is because we share a mental architecture. Without a mental architecture, and an experience of living with such an architecture, we cannot predict what mental processes lie behind those patterns of gaze. Without architecture we could not experience cognitive presence. Without architecture no coherent behavior would emerge.

Architecture

The particular architecture chosen for controlling an NPC is key in determining expandability, modifiability and capability of an intelligent system. Even more importantly, it will greatly affect the kinds of capabilities that we can embed in the characters it will control. Ymir is a three-layered architecture that I built for controlling characters that can interact in realtime with humans (Thórisson 1996, 1999). Among its premises is that any system that has to act in realtime in complex environments needs to have *priority* and *time* infused deeply into its structure. Ymir proposes three levels of priority into which perceptual, decision and action/planning processes fall. It also addresses turntaking by proposing a separation between control of the process of dialogue and the process of content generation and understanding. Through these constructs it affords a highly expandable, broad approach to building autonomous agents.

In earlier work (Thórisson 2005) I proposed four categories of presence cues, based on four distinct kinds of cognitive processing:

- *Reactive cues for cognitive presence*: Behaviors that relate to immediate events or stimuli, environmental or mental.
- *Symbolic cognitive presence cues*: Behaviors that require a skill for manipulating symbols, for language-like skills, production and understanding.
- *Planning cognitive presence cues*: Behaviors expressing an ability to look ahead and use acquired information to predict the future.
- *Holistic cognitive presence cues*: Indications of the ability to coordinate behaviors in the other three categories of presence cues.

Ymir can address all of these, but where it goes beyond most other architectures is with holistic cognitive presence cues: Because of its priority and coordination constructs, it can accommodate decisions at multiple levels of

communicative act, respectively, whether verbal or non-verbal.

[6] The performance of well-rehearsed material may on first glance seem to break this limitation, as actors can pay attention and react to their fellow actors' speech while speaking their own part. This, however, is a good example in support of the present point: The actors are not generating the content of their speech dynamically. If they were, they would not be able to speak and listen at the same time. Moreover, the possibility would become even more remote if their fellow actors were generating new content from scratch. It is *content interpretation* and *dynamic content generation* makes turntaking necessary. For a more detailed discussion of this issue see Thórisson (2004).

detail and timescales, yet produce coherent, believable multi-modal, multi-DoF behaviors in realtime. A prerequisite to the performance is providing the right perceptual information to the various processes in the architecture in time to produce the necessary decisions for responding to them.

We take the concept of *cognitive validity* (V_c) of an architecture to be its potential to do things – perceive, think and act – in the same way that natural cognitive systems do them. If we define "faking it" as the method of producing presence in a system without a valid underlying cognitive architecture,[7] i.e. $V_c = 0$, it can be reasonably deduced from the discussion so far that presence cues in the Planning and Symbolic categories may be more difficult to produce in an artificial system than the Reactive category because (a) they require significant processing power and knowledge represented to work correctly, and (b) they are probably harder to "fake" than Reactive category processes through the use of simple rules or constructs (see e.g. the Loebner Prize[8]). It may also be argued that Planning-type processes have come further in A.I. research than systems producing language – that is, robots are navigating better than they are speaking. This, however, says nothing of whether one is easier to fake than the other; because the goal of A.I. is not cognitive presence and since we don't know the relationship between cognitive validity and cognitive presence, it is difficult to make any such claim. The Holistic category is probably the least studied of the four classes. Because it concerns the integration and interaction of cues from the other categories, it may well be that a closer scrutiny of this category presents the most important category for achieving strong cognitive presence. Holistic presence cues will most likely be the most difficult to implement, because by definition they depend on the correct operation of behaviors in the other categories.

The cognitive validity (V_c) of a system and the strength of the presence (P_s) it expresses could have several kinds of relationships. A direct linear relationship between V_c and P_s provides a strong reason to look closely at cognitive architecture when designing interactive characters. Observed results with simulated humanoids indicate that if cognitive skills and behaviors from the Reactive category are included in an otherwise fairly simple agent, presence is almost certain to emerge (Thórisson 1996). Further, it seems that its strength may be in some ways correlated with the validity of the agent's cognitive architecture.

Now, a note to the skeptical, some of who may ask whether we need a complicated architecture at all – couldn't we simply use recorded data from actual humans interacting to produce the desired features of communication and interaction in graphical characters. After all, motion capture has worked exceedingly well for a lot of movements such as walking, fighting, and a host of other gestures. As those who have worked with motion capture know, there are significant problems using it for interaction with a dynamic environment. One solution is to augment it with other methods, but making seamless transitions between very different motion techniques may then become an even greater challenge. It is not feasible to store responses to all possible scenarios as static sequences. Nowhere is this more obvious than in the human eye: Its movements are along two axes,[9] yet they reveal so much about the mental and physical state of their owner that even a delay of less than half a second can break down or severely put a damper on cognitive presence. Motion capture becomes ludicrous in the case of eyes – it is of no use to stored sequences of eye movements relative to the head, when what really contributes to where we look is the environment (another person we want to look at, a driving car we want to follow with our gaze). The same applies to *all other movements that are made relative to an external, dynamic environment*. This is why face-to-face communication is the "final frontier" in computer games, and why it is so challenging. This is also why we need to pay close attention to the computational architectures we use to

[7] Notice that cognitive presence does not presuppose a cognitive architecture – it is strictly a measurement of the experience of an observer/interactor of some entity.

[8] http://www.loebner.net/Prizef/loebner-prize.html

[9] We can ignore the small rotations human eyes can make around the axis of sight, as this is generally not perceivable by observers.

produce the behaviors.[10] Another often-ignored issue is highlighted via the eyes: To produce a believable pattern of fixations and saccades the eye needs information about its surroundings. This is where the outer feedback loop from Figure 1 comes into play: Without a perception-action feedback loop our autonomous agent will behave like a wind-up puppet, no matter how sophisticated its internal machinery, because it cannot act on the information that lies right before its eyes.

Avatar Intelligence Infusion

Now it's time to put this all in context. The first point I want to make is that as exotic and utopic as autonomous characters with cognitively valid A.I. architectures may sound, there is nothing about what has been said that predicts an all-or-nothing effect: Cognitive validity aside, incremental construction of ever-more-flexible and powerful architectures will move us increasingly closer to full cognitive presence. One reason for taking a serious look at cognitive presence and validity has not to do with autonomous characters per se, but their partially autonomous cousins controlled at a high level by human players (Figure 6). The sheer complexity of the interface connecting human computer game players to their graphically represented avatars makes cognitive presence an important high-level evaluation measure – it provides a way not to get lost in the woods.

Interactive cognitive presence is a lot harder to achieve than the other kinds of presence because interaction requires the system to have an active perception-action loop, and these are difficult to implement well and can be quite complicated to maintain as the system grows. No matter whether we are building NPCs or avatars to be controlled by our human players, we must not ignore the perception-action loop (f_{Ei}, Figure 6). It is this loop that allows living things to react to their environment; it is this loop that separates blind execution of pre-stored sequences from truly reactive and thinking systems.

Cognitive validity as a concept is important to the extent that it helps achieve cognitive presence in the construction of systems; humanoid gaze, for instance, that uses cognitively believable mechanisms for deciding the next fixation, is more likely to produce a sense of cognitive presence than is gaze based on statistical data (a relative of motion capture) from human psychological experiments, as sometimes suggested. In fact, statistics or observational data about how often gaze tends to fall on the face of a person we converse with can only be a guide, as the frequency of certain fixations cannot ever be the *cause* of those fixations – only the (emergent) result of the controlling mechanisms. It is the *mechanisms* that are interesting – not the resulting statistics of what can be averaged over those mechanisms operation. Therefore, cognitive validity cannot be ignored.

Humans use prior experience to judge the strength of the presence; for a simulated human we will get embodied cognitive presence only if the behavior of the virtual human resembles that of a real human in some critical ways. For a given period of behavior, the strength of the perceived embodied cognitive presence will thus be (roughly speaking) a function of (a) the amount of opportunity for the simulated human to express the results of its thoughts through its behavior, and (b) the similarity of its behavior to the perceiver's experience of real humans. As such, it is (typically) easy to recognize and classify in embodied intelligent systems like those we are familiar with: animals and fellow humans. Yet cognitive presence is a phenomenon that is hard to quantify and we may want to develop means to measure it reliably.

At this point it is not clear how, what kind – or whether – a strong sense of human-like cognitive presence will emerge from half-finished or partly accurate cognitive architectures. It is fairly obvious from watching hand-drawn animations that cognitive presence can be achieved in artificial entities. However, animations are non-interactive and created through an elaborate, iterative process. We do

[10] This is true even where the goal is not behavioral realism, although in that case one would be less concerned with cognitive validity and more with cognitive presence. In either case the architectural complexity for controlling autonomous characters is a significant challenge.

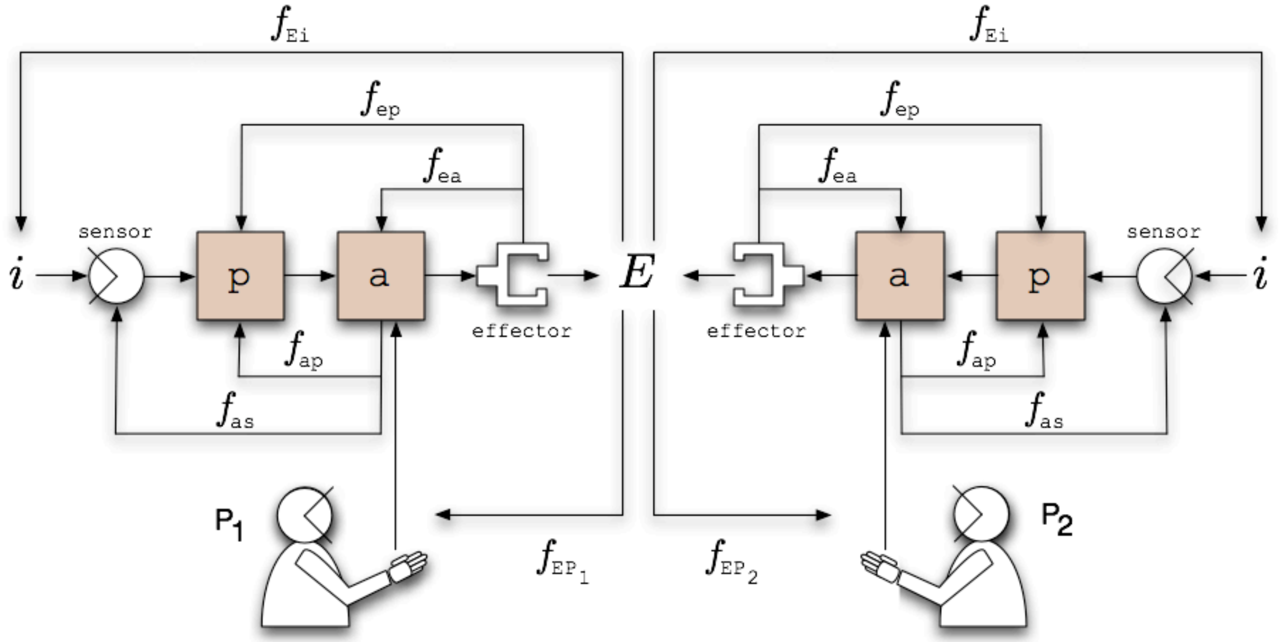


Figure 6. *Two human game players (P1, P2) perceiving an ongoing game environment (E) and controlling semi-autonomous avatars via high-level commands to the avatars' action (a) functions (control interfaces not shown). Both human and machine have multiple feedback loops for taking into account various ongoing processes in their future actions (see Figure 3).*

not have a clear idea of how hard it is to make those entities interactive without loosing this important quality. Using cognitive presence as a guiding light in building computational models for controlling avatars may help game developers achieve a stronger sense of presence – and achieve more engagement – in their game play.

Conclusion

What does a subjective concept like cognitive presence have to do with building avatars and characters with artificial intelligence? A lot, as can be gleaned from the previous discussion. In the early days of any scientific field the scientists' intuition and perception play a key role in advancing the field. In the formative decades simple ideas by a handful of people tend to play a significantly larger role than after the field has matured. The Wright brothers did not sit and work out equations in aerodynamics to get their airplane to fly – in fact, they couldn't have, because the necessary math hadn't even been invented.

The "art" in artificial intelligence is in many ways much more important than many people would like to believe. Defined broadly, it is important from the perspective of human creativity: Creative insight provides a driving force for every science. Like symmetry and beauty sometimes providing mathematicians the route to proper explanation, cognitive presence can serve as a subjective litmus test for intelligence. It is not unlike using our vision to judge the realism of computer-generated dinosaurs and humans in movies. The role of artists in discerning subtle differences between the myriad of alternative methods for driving artificial characters will help scientists choose architectures and select between complex mechanisms that, when looked at by other means, may look alike. In short, a high-level concept like cognitive presence may help scientists and practitioners develop ways of evaluating their creations "at-a-glance". The possibly also exists to make the concept quantifiable with the help of measurement techniques similar to the IQ and personality tests used for measuring high-level (and often quite hypothetical) features of human cognitive function (cf. Sas & O'Hare 2003). This could help game developers who do not care as much about the underlying mechanisms as they do about the final result.

I have claimed that fully autonomous NPCs and avatars controlled by human players will necessarily have to share important parts of a similar control architecture, as the human players will not be able to control the vast number of parameters necessary for achieving high levels of cognitive presence. Further, proper automation – in essence co-piloting of an avatar by human and machine – is likely to benefit greatly from building on top of architectures created for fully autonomous characters. I believe that the material presented in this paper supports the argument that in both cases the controlling architecture itself is a major determinant of the resulting avatar behavior.

In conclusion, there is thus no question that graphics will play an important role in pushing A.I. forward. Conversely, there is no doubt that A.I. will increasingly play a role in pushing the envelope of computer graphics. I have pointed out one such area, avatars and autonomous game agents, but the opportunities go *far* beyond games.

Acknowledgments

I am deeply indebted to my colleague and friend Hannes Högni Vilhjálmsson whose ideas and work has directly influenced the present work and whose close collaboration continues to mold the ideas presented. Thanks to Guðný R. Jónsdóttir, Eric Nivel and Hrafn Th. Thórisson for many excellent comments on the paper. My thanks to Torfi Frans Ólafsson and our collaborators at CCP Inc. Lastly a big thanks to my wife Katrín Elvarsdóttir for patience and support. This work was funded in part by a grant from RANNÍS, and by a Marie Curie European Reintegration Grant within the 6th European Community Framework Programme.

References

Argyle, M. & M. Cook (1976). *Gaze and Mutual Gaze*. England: Cambridge University Press.

Card, S. K., T. P. Moran & A. Newell (1986). The Model Human Processor: An Engineering Model of Human Performance. In K. R. Boff, L. Kaufman, & J. P. Thomas (eds.), *Handbook of Human Perception Vol. II*. New York, New York: John Wiley and Sons.

Cassell, J. & Vilhjálmsson, H. H. (1999). Fully Embodied Conversational Avatars: Making Communicative Behavior Autonomous. *Autonomous Agents & Multi-Agent Systems*, **2**(1):45-64.

Duncan, S. Jr. (1972). Some Signals and Rules for Taking Speaking Turns in Conversations. *Journal of Personality and Social Psychology*, **23**(2), 283-292.

Goldberg, K. (2000). Introduction: The Unique Phenomenon of a Distance. In K. Goldberg (ed.), *The Robot in the Garden: Telerobotics and Telepistemology in the Age of the Internet*, 2-22. Cambridge, MA: MIT Press.

Goodwin, C. *Conversational Organization: Interaction Between Speakers and Hearers*. New York, NY: Academic Press, 1981.

Harnad, S. (1991). Other Bodies, Other Minds: A Machine Incarnation of an Old Philosophical Problem. *Minds & Machines* **1**:43–54.

Hayes, P., K. Ford (1995). Turing Test Considered Harmful. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence,* Montreal, 972-7.

Held, R. M. & Durlach, N. I. (1992). Telepresence. *Presence: Teleoperators and Virtual Environments*, **1**(1), 109-112.

Kleinke, C. (1986). Gaze and Eye Contact: A Research Review. *Psychological Bulletin*, **100**(1), 78-100.

Luce, R. D. (1986). *Response times*. New York: Oxford University Press.

Riesberg, D., Scheiber, R. & Potemken, L. Eye Position and the Control of Auditory Attention. *Journal of Experimental Psychology: Human Perception and Performance*, 7(2), 318-323, 1981.

Sas, C. & G. M. P. O'Hare (2003). Presence Equation: An Investigation Into Cognitive Factors Underlying Presence. *Presence: Teleoperators and Virtual Environments*, **12**(5), 523-537.

Sheridan, T. (1992). *Telerobotics, Automation, and Human Supervisory Control*. MIT Press, Cambridge, Massachusetts.

Thórisson, K. R. (2007). Modeling Multimodal Communication. To be published in I. Wachsmuth & G. Knoblich (eds.), *Modeling Communication with Robots and Virtual Humans*, Springer Lecture Series in Computer Science. New York: Springer.

Thórisson, K. R. (2005). On the Nature of Presence. *AISB 2005 Symposium, Presence Cues for Virtual Humanoids,* Hertfordshire, UK, 12-13 April.

Thórisson, K. R. (2002). Natural Turn-Taking Needs No Manual: A Computational Model, From Perception to Action. In B. Granström, D. House, I. Karlsson (eds.), *Multimodality in Language and Speech Systems*, 173-207. Dodrecht, The Netherlands: Kluwer Academic Publishers.

Thórisson, K. R. (1999). Mind Model for Communicative Creatures and Humanoids, *International Journal of Applied Artificial Intelligence*, **13**(4-5):449-486.

Thórisson, K. R. (1996). Communicative Humanoids: A Computational Model of Psychosocial Dialogue Skills. Ph.D. Thesis, Massachusetts Institute of Technology, U.S.A.

Turing, A. Computing Machinery and Intelligence. *Mind,* **59**(236), 433-60, 1950.

Vilhjálmsson, H. H. (2004). Animating Conversation in Online Games. In M. Rauterberg (ed.), *Entertainment Computing ICEC 2004*, Lecture Notes in Computer Science, **3166**:139-150. Springer.