Designing Humanoid Agents: Some High-Level Issues

11.

Having designed a dialogue-capable agent, and evaluated three versions of this humanoid, it is now time to take a step back and look at the issue of humanoid agent design in a larger perspective. In the real world, physics—and the workings of natural design and selection—dictate the way multimodal creatures look. In the digital world there is much more flexibility (or so we'd like to think). This flexibility (along with the fact that computer characters are neither animals nor human animals) leads us naturally to ask *"What* is my agent?", or more specifically, *"What kind of creature is this (collection of) software?"* The question is important for anyone who wishes to converse with an agent face-to-face.

In this chapter we will look at the question of system validity: What flexibility does the designer of a conversational computer controlled agent have in his designs? How much of that flexibility is limited by human communication capabilities on the one hand and technological limitations on the other? Is the *appearance* and *behavior* of the agent a *valid* representation of its capabilities? The discussion will have more questions than answers, but then again, this is uncharted territory and asking the right questions is more important at this point than providing what would inevitably be wrong answers. Some of the topics touched on here may be a precursor to the systematic evaluation of multimodal communicative systems.

11.1 Validity Types

The validity of a model is determined by reference to the real-world system or object it is intended to be a model of. In the case of communicative humanoids this object would be a human being, or more

specifically, a person engaged in face-to-face interaction. We can distinguish between at least three kinds of system design validity:

1. *Face* validity,
2. *Functional* validity (of control structures) and
3. *Structural* validity.

The first two validity types are especially relevant to the design of computer characters.

11.1.1 Face Validity

Achieving system face validity for a design involves making the system, on the surface, look and behave like the real thing. In the case of dialogue, an agent would have to seem to an observer like the real person engaging in a face-to-face interaction. No questions are asked of the underlying control structures to achieve the observed behavior of the agent.

11.1.2 Functional Validity

The functional validity of a system's control structures is the validity of what that system's control structures *do,* compared to those of the object modelled. For the agent's control structures to be *functionally* valid, the mechanisms controlling its behavior would have to have corollaries in the human mind, at least metaphorically. For example, in the functioning of the agent's system there would be processes and states that we could metaphorically refer to as "thinking", "listening", "attentive", "confused", etc. These processes and states also have a relationship to the system's other components that is functionally equivalent to the way components interact in the real system. A functionally valid system will, for all practical purposes, behave like the system it is trying to model: a response from it will be met with the appropriate counter-response, which in turn will be met with an appropriate counter-counter-reaction, etc. A moment's reflection quickly leads us to see that it is probably very hard to achieve face validity without system functional validity, although theoretically it may not be impossible.[1]

1. An example of face validity being achieved without functional validity is the use of fractal geometry to render very realistic-looking objects such as mountains, clouds, etc. The functional model of these phenomena would be achieved by actually modelling the individual atoms and light rays scattering off these.

11.1.3 Structural Validity

Finally, system *structural validity* is the amount of direct structural correspondence between the model and the object modelled. For a model to be structurally valid, it has to include all the necessary components of its real-world counterpart, including system architecture and its physical properties. For example, if we want to model a mind, it may be necessary to model the structure of the brain, although many believe this will not be required.

Eventually we might want our agents to share the same mental structures as people, but that may not be—and hopefully isn't—necessary for the purpose of building a useful agent. A useful agent, i.e. one that can participate with some skill in dialogue, needs high face validity, but as we already mentioned, this is hard to achieve without at least some functional validity to back it up.

11.2 Functional Validity in Humanoid Computer Characters

Functional validity may be applied to the design of humanoid agents along the following lines: A humanoid agent's outward behavior has to match the user's pre-conditioned (learned) expectation about the relationship between internal processing and morphology of a dialogue participant's behavior (Figure 11-1). For example, the user's mental model of a facial expression's meaning has to match the actual meaning of that facial expression. How to achieve this is an empirical question, and one that is likely to vary between cultures.

Let's put this in a more formal framework. For an agent to be a functionally valid conversant, two pairings that have to happen. The first being the match between the internal state of the machine and its expressions and behavior, denoted $\{\sigma, \sigma'\}$, the second being between the user's recognition of the expression and his or her interpretation of its meaning—i.e. relation to the expressee's mental state, denoted $\{\sigma', \psi\}$. As long as $\{\sigma, \sigma'\}$ and $\{\sigma', \psi\}$ approach a functional[2] correspondence, that is, we can make the assumption that a correct match exists for most or all $\{\sigma, \psi\}$ pairs, the agent's behavior will be a facilitator to the dialogue. In fact, as long as there is a better-than-chance correspon-

2. The term "functional" here is used in the conventional meaning of the term—i.e. what the system *does*. In this view a metaphor from human psychology can plausibly be mapped to the internal workings of the computer, as e.g. "thinking" could correspond to "processing utterance" and "confused" could correspond to "incomplete parse".

dence between the internal state and the expressions, the expressions will eventually always facilitate the dialogue because given time, the person would learn the correspondence. For practical purposes, however, one would want a much-better-than-chance correspondence to avoid frustrating the user.

11.3 What is my Agent?

A central question for agent design over the next decades is how to get around technological limitations that prevent us from achieving functional and face validity. Anyone who has read the preceding chapters in this thesis will by now have realized that physical makeup plays a part in making or breaking the fluidity and naturalness of face-to-face dialogue. And anyone who has ever walked into the bar in Star Wars I knows how hard it is to strike up a conversation with an alien that looks like that in Figure 11-2.

Nevertheless, computer characters should be represented outward in a way that conveys their functionality succinctly, without evoking false expectations in the user. For example, agents equipped with today's computer vision couldn't possibly recognize more than a handful of everyday objects, yet users might mistakenly assume that it can "see"

FIGURE 11-1. The state of the underlying mechanisms (σ) produces a facial expression σ′, which has to match, at least functionally, the user's intuition about the relationship between facial and mental (Ψ) states (σ′ ≡ Ψ). Of course, for a machine this would be a metaphor, and the only measure of its "correctness" is that it is beneficial for the communication.

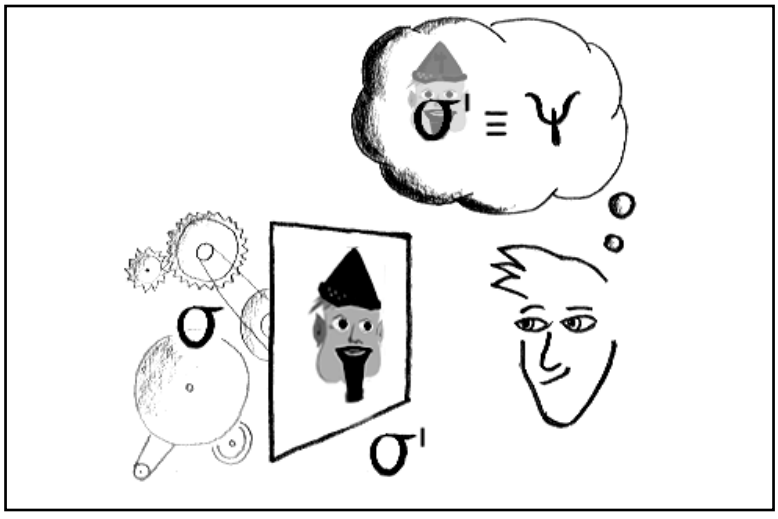

FIGURE 11-2. Anyone who wanders into the Star Wars [1977] bar is likely to wonder how to strike up a conversation with another guest.

objects in the surrounding just like they themselves can. This is a question of misjudged {1} perceptual capabilities. Speaker-independent speech recognition will undoubtedly be limited to a few hundred words for years to come (although pragmatic constraints will enable more top-down processing to improve it as we move toward situated characters). Users of speech-recognizing computers are invariably found to think that computers have a larger vocabulary than they actually do; this is a question of misjudged {2} language capabilities. Giving computer agents human bodies, users may easily think their agility equals their own. This is an issue of misjudged {3} motor skills. And finally, looking human makes people think you are as *smart* as humans. This is misjudged intelligence or {4} mental capabilities.

I think the solution to such inescapable problems lies in clever design; design that clearly shows the connection between appearance and abilities. Even more important is to implicitly and explicitly provide people with indicators of "mental" (computational) limitations in the way the agent *behaves*. An explicit way to achieve this is for example giving an agent the ability to guide a user in her interactions with it (e.g. "I know the names of all the planets but not their moons"). This will prove to be a very efficient guidance tool for helping users to adapt more quickly to the agent's limitations. Giving the agent a reduced ability to speak ("I know planets; not moons") will foster expectations on the user's side that the agent's understanding of speech is similarly limited. An implicit way to provide guideposts to the capabilities of characters is to link their inner functioning to subtle aspects of facial movement, intonational patterns and gaze control.

11.4 The Distributed Agent

As systems become more distributed and options for various kinds of implementations of humanoids become increasingly varied, the following question will eventually come up: Where is my Agent? The issue is more involved than one may think at first sight. This is not a matter of the Agent getting "lost" in cyberspace and cannot be solved by implementing a "search" program to locate your Agent. It is a matter of knowing "where" someone is when you're talking to them. Take a look at the following story.

11.4.1 Where is my Agent?

You're talking to a Mr. Lien at a restaurant. As he sits there in front of you at the small table in your dark little corner, his frontal brain lobes are contained in a jar in his living room at home; his visual processes in an automobile somewhere in Iceland, remotely connected to the two shiny cameras pointing at you form the other side of the table; his body has four arms, but only one of them is visible to you, the rest are plugged into material transporters—transporting his hands to god knows where in the world; and his "legs"—if you could call them that—don't look like they will ever be able to support the rest of him. And you wonder, *"Where* is this character?"[3] Is he at home, where his higher mental functioning resides, is he where is perception his located—somewhere in Iceland, is he where his hands are—where the "action" is, or is he here, the place where all of these seem to meet? You try to answer the question and then you give up. You tell him that you will consider talking to him again when he's pulled himself together. As you walk out you cannot but curse the designers of this agent for being so inconsiderate of its user's communicative needs.

And why did you get frustrated? For one, because Mr. A. Lien is an alien, you couldn't predict how it would respond to your actions, how its memory worked, or how it perceived you and the environment you met in (we covered this in the last section). Second, any references to events in the immediate environment such as actions of A. Lien during the conversation and the waitress that brought you the Brainblaster cocktail, were precluded because you had no way of knowing whether A. Lien had been paying attention (whatever that means for an alien) during those events. Since there is no centralized, localized place which action and perception are limited to (via a body), he could have been doing anything anywhere and not been present at all (a pair of cameras prove nothing). But most importantly, because there was a communica-

3. Readers of philosophy may recognize the theme of this muse—its precursor can be found in Dennett's [1981] excellent short story "Where Am I?"

tion time lag between the wrong pieces of Mr. Lien's "brain", its back channel never matched your pauses, the movement of the cameras was out of sync with its verbal output, and when it fell silent every now and then it was impossible to predict whether this was due to a "mental processing" delay, a "brain communication" delay or a general failure of a significant piece of its mental machinery.

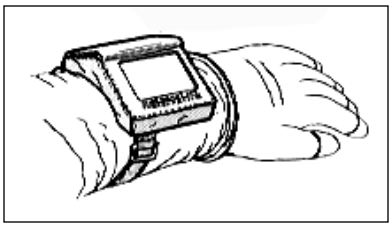

FIGURE 11-3. NASA scientists have designed a wrist-based computer with a touch-screen, intended as a portable assistant on space walks [NASA Tech Briefs 1995b].

11.4.2 Wristcomputer Humanoids

We will consider one version of the mobile agent: an agent that appears in the display of your watch. This idea isn't pure fantasy or science fiction; NASA has recently done some research with wrist-based, touch sensitive screens for space walks (Figure 11-3), AT&T are doing research on wristphones and various research is making it possible to miniaturize communications technologies, displays [Depp & Howard 1995], cameras and CPUs (Figure 11-3).

The question here is: what metaphors do we want for communication with machines that can have a distributed "brain" and "body"? If my machine agent talks through my watch, but has no visual sensory apparatus to sense me (or anything surrounding my watch) is the right metaphor that my agent is "in" my watch, or is it more accurate to talk about it being somewhere else, talking to me through a visual walkie-talkie? How about the situation where there is some sensory apparatus in my watch, but only very little? How about if all the sensory apparati were in my watch, but it's brain is somewhere else? These questions revolve around three things: {1} bandwidth limitations, {2} the breakup of an agent's mental functioning and {3} the metaphors we choose for the communication.

I will try to argue that the answers to these questions should be based on the mental limitations of the human user, and the limitations of the metaphors we use to simplify the interaction, *the communicative humanoid*.

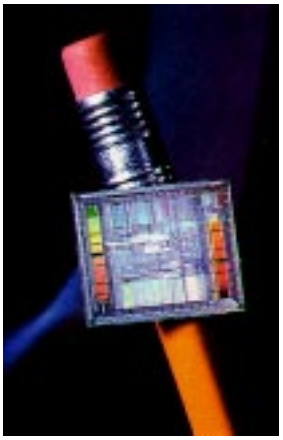

FIGURE 11-4. The Alpha 21064 chip from Digital Equipment Corporation measures 1.39 x 1.68 cm in size (this picture shows it roughly two times the actual size) and contains 1.68 million transistors. This is the chip that runs most of Ymir.

FIGURE 11-5. When humanoids start appearing on the LCD screen in our wrist computers, how will the primary communication problems manifest themselves?

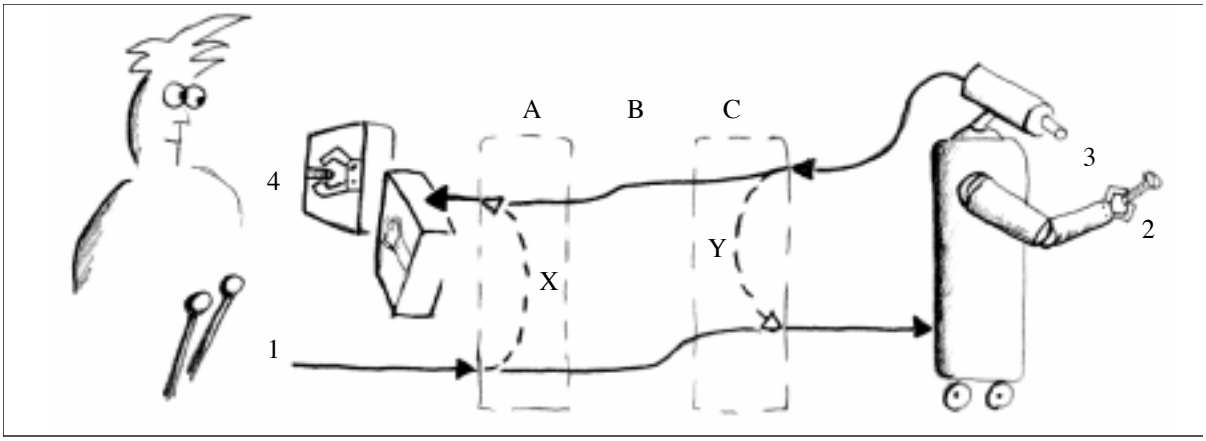

FIGURE 11-6. In a telerobot supervisory system, several paths define information flow. Here, the main loop of information flow is that from controls [1] to the robots actuators [2], through the robot's sensors [3] and back to the operator's displays [4]. System A provides a local feedback to the operator [X], system C provides a local feedback loop within the telerobot [Y]. Gap B represents some barrier, time, distance or inconvenience. (After Sheridan [1992].)

The question we need to answer is *Where can we accept limited bandwidth in multimodal communication?*

11.4.3 A Comparison to Teleoperation

There are several corollaries between teleoperation and agents with "distributed psychology". In Figure Figure 11-6 three loops characterize information flow: {1} The flow form controls to robot and back to displays, {2} the flow from controls directly to displays, and {3} the flow from robot sensors to robot manipulators. The reason these loops are important is the fact that a barrier exists in the transmission channel between the supervisor and the robot. This provides the reason for a local loop; a local loop displays immediate (time specific) data about the manipulation of controls without going through a sub-optimal channel, thus increasing the rate at which the operator can update his model of his own actions. Figure 11-7 shows the situation in face-to-face dialogue: Here the local loop goes through reactive paths in the robot itself [x]. This loop is responsible for responses under 1 second. A higher-level, slower loop takes care of administration of responses related to the process of dialogue. The third loop represents data flow through the rest of the agent's knowledge and reasoning systems. This is the slowest loop. The reader may recognize here the gross anatomy of Ymir. These loops all have a fixed relationship to one another (see Chapter 7.,

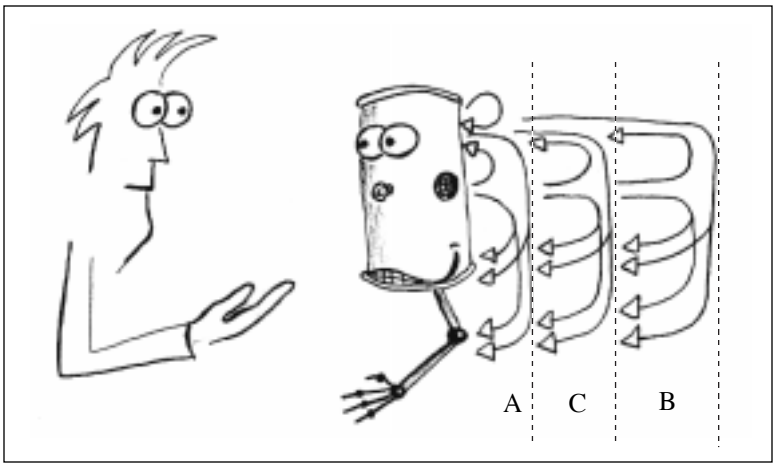

FIGURE 11-7. Many feedback loops exist in dialogue systems. A rough comparison between telerobotics and face-to-face conversation reveals some structural similarities. Here the human takes the role of the teleoperator, while the tincan humanoid corresponds to the role of a telerobot. Sections A, B and C in this figure correspond roughly to sections A, B and C in Figure 11-6. Loop A is highly reactive and the robot has very little control over it, yet it is to great benefit to the "operator" (human). (Of course such tight control loops also benefit the robot, e.g. the loop from the eye input to the eye's muscles). Loop C is the local loop of the robot—its "inner agenda". Loop B is the barrier that may exist with the input from the human to the robot's knowledge—this loop may be characterized in the same way as loop B in Figure 11-6, it constitutes time, distance and/or inconvenience.

page 92) which needs to be maintained in order for the dialogue to proceed at normal speed. Both the lowest and middle loop are highly time-specific; the top loop is semi-independent of time.

As we mentioned in "Back-Channel Feedback" on page 40, and as shown in the human subjects experiment ("Human Subjects Experiment" on page 161), breaking the lowest loop may result in discontinuities in the dialogue and overall lower satisfaction with the interaction. For an agent situated in a watch, with limited computational capacity in the device itself, it may be tempting to leave only the sensing devices in the watch, and move all computationally intensive processes to a remote location. However, as teleoperation has shown, this may lead to aliasing effects where the operator moves faster than the actual data display rate allows for. This invariably leads to error in operation, the most obvious in dialogue being overlaps in speech. All delay constants in the system (which are in fact unlikely to be constants) have to be measured to guarantee timely execution of events in each loop. To ensure the correct update time for gaze, its movements need to be driven by informa-

tion flowing in a tight loop from eye input to the gaze controlling mechanism. The update time for back channel feedback is (mainly) achieved by a close loop from the agent's hearing mechanism, to the vocal motor control and head motion.

The above analysis can be used to determine where we can "chop up" the wrist computer agent's psyche to distribute its functioning and make better use of computational resources. The reader shouldn't be surprised that what emerges is the basic structure of Ymir. For the loops in Figure 11-7 we have A = 100 [0 ~ 250] msec, C = 250 [150 ~ 1000] msec, and B > 1000 msec. Here we have approximate average total transmission times for each part of these loops. Notice that this is not just data transmission time, it is the *complete loop time*—data collection, processing, decision making, motor composition and motor execution. Thus, even with a very high bandwidth between the wrist computer and the remote location, going through a geosynchronous communications satellite you will always introduce a transmission delay of not less than 200 ms (uplink-downlink). That is clearly too high for loop A and maybe for loop B as well. A cellular connection will serve us better, but loop A would probably still need to be local to the agent's display. Several studies have indicated that this is precisely the reason why videophones and video conferencing hasn't caught on as was expected when the development of this technology started in the '50s [Whittaker 1994, Whittaker & O'Connaill 1993]: because the technology cannot support the high bandwidth necessary for correct synchronization between image and sound, as well as uneven refresh rates, the feedback in the reactive loop gets lost in the process. This leads people to choose telephones over videophones, where there is a higher data transmission rate and less synchronization problems.

11.5 Conclusion

Undoubtedly many problems will come up as we design more sophisticated agents and the systems get bigger and more complex. One problem with copying an activity such as face-to-face interaction in a machine, that integrates perception, planning and action, is scaling. This cannot be approached like a telephone network, where the mathematics of adding a certain number of new users is well understood and the problem scales well. The importance of using guidelines such as those presented in this chapter in pinpointing where possible problems could arise, and what those problems might look like, cannot be underestimated. However, this is just the beginning: We need to go far beyond the current understanding of communication and telerobotics to

be able to accurately estimate the efficiency, problems and satisfaction with such systems. But before we know how to design the systems, perhaps it is too soon to try to design evaluation methods.