Ymir / Gandalf: An Evaluation in Three Parts

10.

In this chapter we will be asking several questions. Having spent all this effort designing and implementing a computer controlled character, a key question is, *Does this system really behave like a human in a conversation?* The answer to that is "yes and no": On the one hand, people seem to give Gandalf a very favorable rating in comparison to humans. (For example, on a scale from 0 to 10 for language capabilities, humans getting a perfect 10, naïve computer users gave the humanoids a mean score of 7.59; SD=1.45.) On the other, the system's limitations are usually obvious to anyone after only 1-3 minutes of interaction. I will answer this question in three ways: {1} By comparing the performance of Ymir/Gandalf to the *Model Human Processor*—a predictive model of human perceptual, motor and cognitive performance [Card et al 1983, 1986], by {2} presenting the results of an experiment with 12 subjects interacting with 3 different characters, and by {3} careful observation from a designer's perspective. The question of concern in the last issue is how easy it is to construct an agent in Ymir. Once the foundation had been laid, it took only between 3 and 5 weeks[1] to construct a minimal set of perceptual, decision and behavior modules for Gandalf. We will look at this at the end of the chapter.

10.1 Evaluating Gandalf with the Model Human Processor

The Model Human Processor [Card et al., 1986; 1983] is a general engineering model of cognition, designed specifically to predict human performance and reaction times. By comparing Gandalf to this model, it

1. This is a rough estimation since some of Gandalf's modules were created in parallel with the development of Ymir.

gives us not only the ability to compare outward behavior of the character to the way humans behave, it allows us to compare the internal components of the Ymir/Gandalf system to something that has proven to be a good predictor of human performance.

The Model Human Processor can be described as a collection of *storages* and *processors* together with a set of *operating principles*. The perceptual system consists of sensors and short-term memories. The cognitive system receives symbolic information from the perceptual system and uses information in long-term memory to make decisions about how to respond. The motor system carries out the responses. Each of these subsystems has a hypothetical processor running at its own clock speed. A number of parameters characterize the behavior of each of these systems. For our situation, the relevant parameters would be:

- τ_c = cycle time of cognitive processor: 70 [30~100] msec
- τ_m = cycle time of motor processor: 70 [25~170] msec
- τ_p = cycle time of perceptual processor: 100 [50~200] msec
- fix = duration of a fixation-saccade pair: 230 [70~700] msec.

The brackets indicate the extremes for a given parameter (*typical value [lower bound*, *upper bound]*, respectively). Card et al. [1986] give a concise explanation of the origin of these numbers. They come from various psychological literature on reaction time and human performance. In this model, all activity in the cognitive system is a result of a discrete number of processing cycles.

For a task such as multimodal dialogue, a person would need to use all of the three subsystems of the MHP, the perceptual system, the cognitive system and the motor system. The sequence of actions in the MHP is *perceive*, *think*, *act*. Thus, cycle times (τ_p, τ_c, and τ_m) occurring within a single step should added; the maximum values for each step are then added together to get a final RT prediction.

10.1.1 Perceptual Processes

For reactive actions, such as perceiving whether the user has stopped speaking, the Model Human Processor would predict 100 [50~200] msec. We could take the lower end of this range to apply to the reactive perceptual processes, using Card et al.s' notational system, analysis of the current implementation of Gandalf shows this to be:

- Read visual (body) data: 8 [5 ~ 10] msec
- Read prosody (intonation) data: < 1 msec
- Read speech (word tokens) data: 2 [0 ~ 5] msec
- Update perceptual processes in Reactive Layer: 8 [5 ~ 10] msec.

These are serial events, giving us *18 [10 ~ 25] msec*. In addition, feeding the above processes with data is done over a fiber-optic net:

- Transmission delay for body data: 10 [3 ~ 30] msec
- Transmission delay for intonation data: 10 [3 ~ 30] msec
- Transmission delay for speech data: 10 [5 ~ 30] msec.

The transmission delays are parallel, giving us 10 [5 ~ 30] msec. This adds up to a total of 28 [13 ~ 55] msec. To answer the question about the speed of determining that a user has stopped speaking, we have to look at the delay from stimulus onset (when the user starts speaking) until the information is available to other processes. In Gandalf, this goes through the intonation tracking system, which has the additional delay:

- Speech on/off filtering: 10 [0 ~ 20] msec
- Silence inertia (constant): 50 [50 ~ 50] msec.

The silence inertia filters out pauses shorter than 50 msec. Taken together, we have *88 [63 ~ 125] msec* to detect that the speaker is silent. For other features, such as detecting that the hands are in gesture space or that the user is looking at the agent, we get a somewhat lower number of *28 [13 ~ 55] msec*. The MHP predicts 100 [50 ~ 200] msec.

The above numbers may seem pretty good, however, no time-dependent perceptual processes have been implemented, and the features detected are relatively simple (for example, no processes are devoted to determining the reliability of the data, which surely must be part of the 100 msec attributed to humans). The only perceptual process at the Process Control level is the deictic-gesture detector, which means that the numbers are likely to be different for a more capable agent.

10.1.2 Cognitive Processes

For an action such as deciding to act on a set of stimuli, the MHP would predict 70 [30~100] msec. Since the Decision Modules only look for logical combinations of conditions to compute their state, the performance for each module surpasses this prediction. Even taken as a group (running on a fast serial machine), the modules take < 0.0 [0.0 ~ 5.0] msec to execute one loop. Adding to that the delay to transmit the decision to the Action Scheduler, which we take to be 10 [3 ~ 30] msec, we get a total of 10 [3 ~ 35] msec. Since no time-dependent decision modules were implemented, it is difficult to predict how this would change for a fully-fledged decision system, although the numbers look like there is room for much more computationally intensive computations.

Adding to the above the time needed to parse incoming speech, update the knowledge bases, and monitor real-world acts, 3 [0 ~ 10] msec, giving a total of *13 [3 ~ 45] msec*.

10.1.3 Motor Processes

For an action such as moving the eyes, the Model Human Processor would predict 230 [70~700] msec. The gaze of Gandalf is updated only about every fourth second, falling short of the observed human performance. However, symbolic gaze events, such as turn signals, are still performed at the right transitions. Other motor responses should fall along the lines of 70 [25~170] msec, according to the MHP. Combining the behavior morphology selection and action scheduling, which is 30 [20 ~ 150] msec, and net transfer, 10 [3 ~ 130] msec, we get *40 [23 ~ 280] msec*. Another limiting factor in the motor system is the performance of the motor system itself, which is close to the upper limit of the Model Human Processor: ToonFace's (Appendix A1, page 203) smallest unit of execution in the current implementation is a constant of *150 msec*.[2]

10.1.4 Full-Loop Actions

The MPH predicts that a "closed-loop" motor task with visual feedback should be limited to 240 [105 ~ 470] msec [Card et al. 1983, p. 35]. Taking together the sequential events in Gandalf, we get

- Perceiving: 28 [13 ~ 55][3] msec
- Deciding: 13 [3 ~ 45] msec
- Acting: 190 [173 ~430] msec
- *TOTAL:* *231 [189 ~ 530] msec*.

These numbers are surprisingly close to those predicted by the MHP. However, a main issue in making a reactive conversant seems not to be reactivity in a closed-loop visuo-motor task, such as the above, but in making the right predictions about where, when and why events happen. In other words, top-down hypotheses must be at work in human-human dialogue to enable turn transitions with 0 msec overlaps in speech [c.f. Sacks et al. 1974], among other things. Another confounding factor is the slowness of the speech recognition, taking between 1.5 - 2 seconds to provide the content of the speech. Gandalf has to do an awful lot of

2. This number was determined empirically for a wide range of motor commands and scheduling loads, and is completely dependent on the speed of the computer responsible for the animation.
3. The lower values were selected since these represent a much larger set of events in Gandalf than the values for speech onset-offset detection.

	Gandalf / Ymir Alpha	MHP
Reactive Perception	28 [13 ~ 55] msec	100 [50 ~ 200] msec
Decision Making	13 [3 ~ 45] msec	70 [30~100] msec
Motor Actions	40 [23 ~ 280] msec	70 [25~170] msec
Full-Loop Actions	231 [189 ~ 530] msec	240 [105 ~ 470] msec

TABLE 10-1. Summary of comparison between Gandalf/Ymir Alpha and predictions of the Model Human Processor.

filling in with nonverbal behaviors to justify to the user this long pause before he responds. We will take a closer look at this in Section 10.2.

10.1.5 Conclusion

The intention here was to compare the current implementation of Gandalf to human performance, as modelled in the Model Human Processor. With the exception of speech recognition, Gandalf's performance stands fairly well up to human performance as predicted by the MPH, for the limited actions it was designed to do. If these performance numbers can be kept when adding increasingly sophisticated processes and modules to the system, one should expect a very reasonable model of a human conversant.

10.2 Human Subjects Experiment

The purpose of this experiment is to evaluate characters constructed in Ymir as they perform in real-time face-to-face interaction with a person, and to evaluate user attitudes toward humanoid interfaces as a function of the type of feedback given by the system. Three prototype humanoids were video-taped in their interaction with the subjects. Subjects' evaluation of the system was subsequently collected with a questionnaire. Subjects' speech patterns and behaviors were scored along the dimensions of relative number of user utterances (number of subject contributions[4] to the discourse over the number of a character's contributions) and relative number of subjects' hesitations and expressions of frustration (over the total number of their contributions).

4. A "contribution" is defined here as any speech utterance that is meant to elicit information, achieve an action, or be a response providing the information or achieving the action.

10.2.1 Background & Motivation

Research intended to answer questions about the various features of agent-oriented systems—that is, systems that employ an embodied, humanoid characters—has to date been hampered by the lack of real computer systems capable of sustaining and supporting spoken dialogue with a human user. To assess topics such as believability, trust, effectiveness of communication, users' likability of the interaction, as well as the question of whether to employ human-like figures to represent the system, these studies have turned to Wizard-of-Oz techniques [Maulsby et al. 1993, Hauptman 1989], mixed automation/Wizard-of-Oz [Thórisson 1992], typed natural language [Neal & Shappiro 1991, Wahlster 1991], iconic embodiments of various types [King & Ohya 1996, c.f. Maes 1994], or simply ignored the issue of embodiment [Sparrell 1994, Thórisson et al. 1992]. As a result, one cannot justifiably generalize the results of these studies and/or systems to future systems employing computer-controlled characters capable of real-time dialogue—tempting as it may have been to many researchers.

Prior efforts have often tried to assess the value of the *very idea of the agent metaphor* using a grab-bag of interaction methods. Because interaction method may be expected to interact strongly with users' perceptions of a system, such methodology is suboptimal at the best, unacceptable at worst. Various research has also intended to evaluate numerous *representations* of agents—humanoid, iconic, animal-like, etc.—by using collections of arbitrary behaviors, or simply ignoring behavior. Instead of trying to evaluate the inherent value of the humanoid agent metaphor, or the value of various visual and auditory representations for computer-imparted agency, I propose to turn these approaches on their heads, using a real computer-controlled humanoid to study communicative behaviors that *require* a humanoid representation.

Since attempts to evaluate full-duplex multimodal systems that employ artificial agents (fully or partially automated) have been virtually nonexistent, no data exists yet on important features of dialogue such as back channel feedback, mixed representations (e.g. spatial gestures + speech), and flexible turn taking, in the natural manner they combine to sustain and support face-to-face dialogue. Yet these are arguably the strongest reasons to employ an embodied, humanoid agent in a co-spatial, co-temporal communications system that uses spoken natural language.

In this experiment, we are interested in features that cannot be reproduced in any other way but by the use of embodied, social actors: spontaneous manual gesture and speech. If creating complex characters with multi-layered input analysis and output generation is to be justified, how else to justify it than with hard data from real interaction? "But why not

compare Gandalf to a condition using no embodiment?" you might ask. One cannot use the conventions of face-to-face dialogue (e.g. "If I'm looking at you while speaking, I'm probably speaking to you") if the conversants are not co-present. "But how about comparing Gandalf to a keyboard condition?" If the goal is to look at *natural speech* and/or *full-duplex multimodal* as interaction, one cannot introduce a keyboard or mouse into the system without compromising its naturalness. This is a different question—one that has been investigated by other researchers [e.g. Seu et al. 1991]—and will not be addressed here.

10.2.2 Goals

One claim that is often heard is that there is no need for multiple modes since the speech channel carries all the necessary data [Ochsman & Chapanis 1974]. If this is true, there is little reason to put in the effort to embody the system, all that is needed is speech synthesis and recognition. The main objective of this experiment is determining the importance of what we refer to here as *envelope* feedback to the effectiveness of dialogue. Envelope feedback includes back channel feedback [Yngve 1970], attentional feedback and other process-related feedback. Included in envelope feedback are reactive behaviors—behaviors that are very quick and people normally don't think about when performing during conversation. Examples include blinking, determining fixation points from moment to moment, saying "aha" at the right times, etc. We also group manual beat gestures in this category. The claim here is that these kinds of behaviors are the strongest argument for using an embodied agent in speech-based human-computer interaction, and, unless shown to somehow be important to dialogue, would be dismissed as yet another useless hog of processor cycle time.

Another kind of feedback that is often mentioned in relation to embodiment are *emotional emblems*. Emotional emblems are facial expressions that reference a particular emotion, without requiring the person showing the expression to feel that emotion at the moment of expression [Ekman 1979]. In the literature on anthropomorphism, emotional emblems have been held again and again to be a feature that an embodied agent-based interface could—and should—add to human-computer interaction [cf. Hasegawa et al. 1995, Nagao & Takeuchi 1994, Takeuchi & Nagao 1993, Britton 1991].

To investigate these questions, we compare three conditions. The basic condition contains content related feedback only. Content feedback is any uni- or multimodal actions that pertain to the topic of the dialogue, such as answers to questions or responses to requests. The second condition adds envelope feedback to the content responses, as defined below. The third condition combines emotional emblems with content responses.

Definitions of Behaviors

The following agent behaviors were used in the experiment:

I. Response to content:

1. Executing commands & answering questions.

II. Emotional emblems:

2. Confused expression when it doesn't understand an utterance.
3. Smiles when addressed by the user and when responding to a multimodal act.

III. Envelope feedback:

— Attentional:

4. Appropriate head turning and deictic gaze when listening to user and executing commands in the domain.

— Back channel:

5. Averting gaze and lifting eyebrows when taking turn.
6. Gazing back at person when giving turn.

— Status:

7. Eye blinks and tapping fingers to show that it is "alive".

— Content-related:

8. Manual gesture when providing verbal content.
9. Verbal acknowledgment when having understood a multimodal act.

We can take behavior 1 as given in any purposeful, communicative system: without appropriate response to content, there is little point to dialogue. But what about the latter two? In an anthropomorphic interface, which is more important: providing the system with the ability to provide (a) emotional emblems, or (b) feedback which is related to the process of the communication? We are claiming that the importance of anthropomorphism lies first and foremost in its power as a unifying concept for simplifying discourse. If this is true, feedback that relates directly to the process of the dialogue should be of utmost importance to both dialogue participants, while any other variables, such as emotional displays, should be secondary.

10.2.3 Experimental Design

Three conditions were tested: The Content Feedback (CONT) condition includes behavior 1 only, thus excluding emotional and envelope feedback. The Envelope Feedback (ENV) condition included all behaviors except II, excluding emotional feedback. The Emotional emblems (EMO) condition includes behaviors I & II, thus excluding envelope feedback. Examples of neutral, smiling and puzzled expressions for each character are given in Figure 10-1.

Hypotheses

Eight hypotheses were tested:

{H1} *No difference will be found for relative contributions from users between conditions CONT than in condition EMO.*

FIGURE 10-1. The three faces used in the experiment. Rows, from top to bottom: Gandalf, Roland, Bilbo. Columns, left to right: neutral expression, puzzlement, and smile.

{H2} *Relatively fewer subject contributions will be found in condition ENV than conditions CONT and EMO.*

{H3} *No difference in hesitations will be found between conditions CONT and EMO.*

{H4} *Relatively fewer hesitations will be found in condition ENV than in conditions CONT and EMO.*

{H5} *No difference in overlaps in speech will be found between conditions CONT and EMO.*

{H6} *Relatively fewer overlaps in speech will be found in condition ENV than in conditions CONT and EMO.*

{H7} *No difference will be found in subjects' rating of the agent between conditions CONT and EMO.*

{H8} *Subjects in condition ENV will rate the agent higher than those in condition CONT and EMO.*

Data for hypotheses 1 and 2 was collected by analyzing video tape recordings of the subjects. Relative number of contributions, as well as hesitations and frustration responses were scored according to pre-determined scoring schemes (Appendix A3, page 215). Data for hypotheses 3 and 4 was collected with questionnaires (Appendix A3.3, page 201).

Variables & Statistical Procedure

The dependent variables of concern are:

1. Relative number of contributions.
2. Relative number of hesitations.
3. Subjects' rating of agent on numerous scales.

The independent variables of concern are:

1. Amount of multimodal feedback (groups ENV, CONT and EMO).
2. Computer character.

The difference between conditions CONT, ENV and EMO on all dependent variables was tested with a repeated-measures MANOVA.

Procedure

Three different characters (face[5] + voice) are used to represent the computer in each condition, each of which was presented equally often in each position, and equally often for each of the conditions:

5. I would like to thank Hannesi Vilhjalmssyni and Roland Paul for designing the faces of Bilbo and Roland, respectively.

Condition		Character	
Content	(CONT)	Gandalf	(G)
Emotional	(EMO)	Bilbo	(B)
Envelope	(ENV)	Roland	(R)

and then varying the order of these conditions for each participant, creating the following presentation order for the 12 subjects, for the three conditions:

Subject	Order of Characters X / Y / Z	Order of Conditions 1st / 2nd / 3rd
1	G / B / R	ENV / CONT / EMO
2	B / R / G	CONT / EMO / ENV
3	R / G / B	EMO / ENV / CONT
4	R / G / B	ENV / CONT / EMO
5	G / B / R	CONT / EMO / ENV
6	B / R / G	EMO / ENV / CONT
7	B / R / G	ENV / CONT / EMO
8	R / G / B	CONT / EMO / ENV
9	G / B / R	EMO / ENV / CONT
10	G / B / R	ENV / CONT / EMO
11	R / G / B	CONT / EMO / ENV
12	B / R / G	EMO / ENV / CONT

The procedure for each subject was as follows:

1. Subject read the instructions for interacting with the characters (Appendix A3.2, page 215), and
2. read and signed a Declaration of Consent. Then they
3. answered a Background Questionnaire.
4. Subject put on the input devices (microphone, jacket & eye tracker) and went through a calibration procedure [Bers 1996].
5. Subject interacted for 2-4 minutes with character X (pilot).
6. Subject interacted with character X for 7-10 minutes and subsequently
7. answered Evaluation Questionnaire for character X (Appendix A3.3, page 216).
8. Subject interacted with character Y and subsequently
9. answered Evaluation Questionnaire for character Y.
10. Subject interacted with character Z and subsequently
11. answered Evaluation Questionnaire for character Z.

12. Subject answered Prior Beliefs Questionnaire (Appendix A3.4, page 219).
13. Subject read debriefing statement.

All three Evaluation Questionnaires are identical, except for the name of the character last interacted with.

Experimenters & Subjects

K.R. Th. acted as experimenter. A convenience sample of 12 volunteers between the ages of 22 and 37, both male and female, were tested. The Background Questionnaire confirmed that the subjects were naïve computer users, with no visual problems or other handicaps. All were native English speakers. Video tapes were scored independently by two scorers, in a double-blind design[6]. Scoring reliability for the variables obtained from the videos, overlaps, hesitations, and number of contributions, was > .95 [Pearson's correlation coefficient, $p < .001$, one-tailed].

Hypothesis	Means	t (pared)	Significance	Conf.
{H1} Contributions: EMO = CONT	EMO=1.52 CONT=1.33	-1.45	n.s. (two-tailed)	√
{H2} Contributions: ENV < CONT, EMO	ENV=1.23 C+E/2=1.42	2.49	$p < .016$ (one-tailed)	√
{H3} Hesitations: EMO = CONT	EMO=0.022 CONT=0.023	.07	n.s. (two-tailed)	√
{H4} Hesitations: ENV < CONT, EMO	ENV=1.0 C+E/2=0.02	-2.86	$p < .008$ (one-tailed)	no
{H5} Overlaps: EMO = CONT	EMO=0.036 CONT=0.015	-1.55	n.s. (two-tailed)	√
{H6} Overlaps: ENV < CONT, EMO	ENV=0.42 C+E/2=0.03	-2.05	$p < .033$ (one-tailed)	no
{H7} Agent Rating (Q1): EMO = CONT	EMO=40.67 CONT=44.83	1.86	n.s. (two-tailed)	√
{H8} Agent Rating (Q1): ENV > CONT, EMO	ENV=46.83 C+E/2=42.75	-3.99	$p < .003$ (one-tailed)	√
{H7} Helpfulness (Q3): EMO = CONT	EMO=3.23 CONT=3.02	-1.13	n.s. (two-tailed)	√
{H8} Helpfulness (Q3): ENV > CONT, EMO	ENV=3.85 C+E/2=3.13	-4.29	$p < .002$ (one-tailed)	√

TABLE 10-2. Results from paired t-tests for each of the hypotheses. Rating hypotheses were tested with two questions from the Evaluation Questionnaire (Appendix A3.3, page 216). DF = 11 for all t-values. EMO and CONT are pooled for all comparisons with ENV. Leftmost column lists which hypotheses were confirmed.

6. My thanks to Katrín Elvarsdóttir and Roland Paul for their precise scoring.

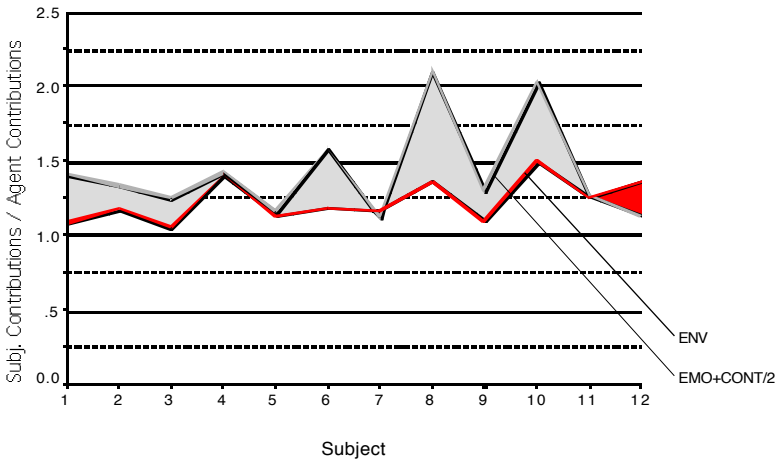
10.2.4 Results

All but two of the eight hypotheses were confirmed. The null hypotheses—that all numbers came from the same pool—was tested with a repeated-measures multiple analysis of variance (MANOVA) with all variables[7], and was rejected [F = 2.742, DF = 24, p < .02] (α is set at .05 for all hypotheses). Overall, the results supported the significance of envelope feedback over emotional emblems and content only feedback: Comparisons between individual means was done with paired t-tests, and are summarized in Table 10-2. No effects were found for order of character [F = 1.86, DF = 6, n.s.] or order of conditions [F = 1.85, DF = 6, n.s.], or interactions between these.

Relative Number of Contributions

Figure 10-3 shows the distribution of relative contributions for each condition in the experiment. This difference was significant at the .01 level [F = 2.74, DF = 10, p < .01, repeated-measures MANOVA]. Figure 10-2 shows a comparison between the number of contributions with CONT and EMO pooled ((CONT+EMO)/2).

FIGURE 10-2. Difference in relative number of contributions for all subjects between condition ENV (dark line) and (CONT+EMO)/2. This difference was significant. The amount of difference between the two conditions is filled with grey; the dark tail at the right indicates a reversal of the overall pattern for subject 12.



7. Set includes Evaluation Question 1, number of hesitations, relative number of contributions, and number of overlaps.

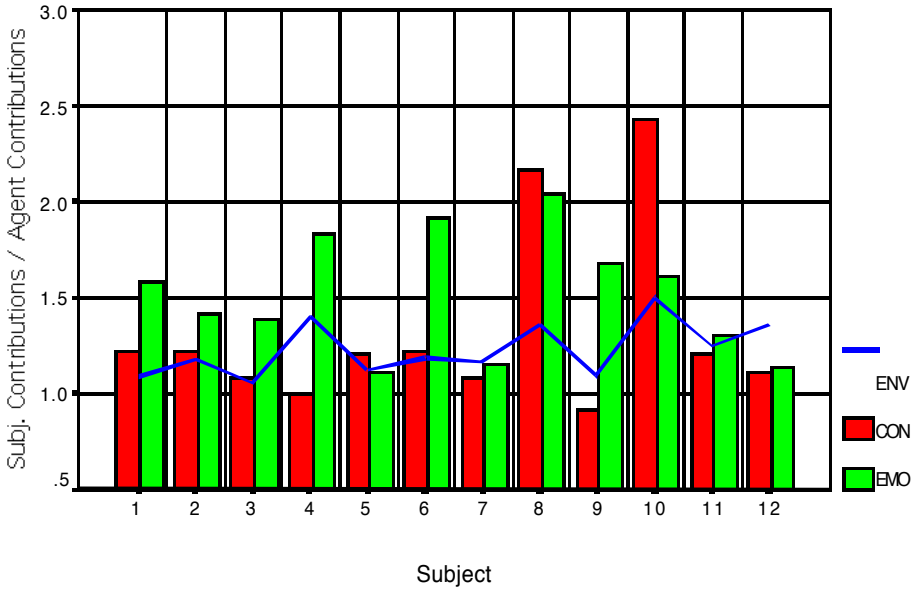


FIGURE 10-3. Relative contributions of subject (Csubject/Cagent) for each of the three conditions (left bar = CONT; right bar = EMO; line = ENV). The difference between the three conditions is significant at the .035 level [F = 2.74, DF = 10, $p < .035$, repeated measures ANOVA]. The difference between EMO and CONT is not significant, however, but the difference between ENV and (CONT+EMO)/2 is (see Table 10-2). In the ENV condition the ratio of user to agent contributions is 1.23.

Figure 10-4 and Figure 10-5 show the means for the two questions that were used to test the hypotheses for subjects' attitudes toward the agents.

Subject Evaluation of Conditions

The subjects' rating of the characters' language abilities are interesting: On a scale from 0 to 10, humans getting a perfect 10, subjects gave agents in the ENV condition a mean of 7.25 (SD=1.86) for language understanding and 7.92 (SD=1.83) for language use. These numbers are surprisingly high, and unless they simply indicate the user's satisfaction with the language part of the system—which, with naïve computer users could be the case—may point to a lack of grounding the lower end of the spectrum (i.e. stating that a dog should get a 1 might have resulted in different numbers).

The sub-questions in question 1 that were significantly different between ENV and the other two (pooled; CE=(CONT+EMO)/2) condi-

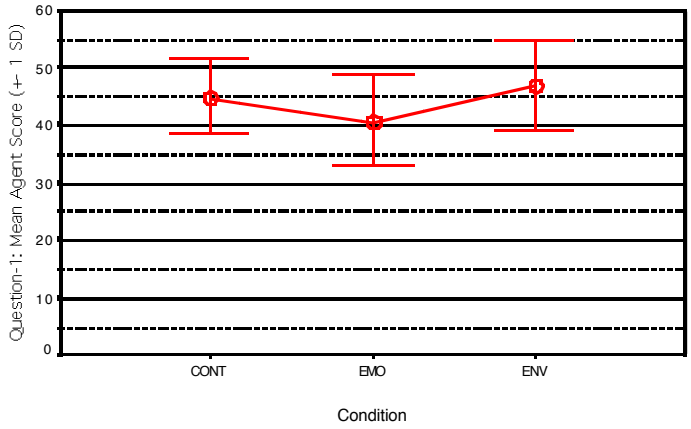


FIGURE 10-4. Means for each of the three conditions for the subjects' rating of the quality of interaction (question 1, Evaluation Questionnaire (page 216).

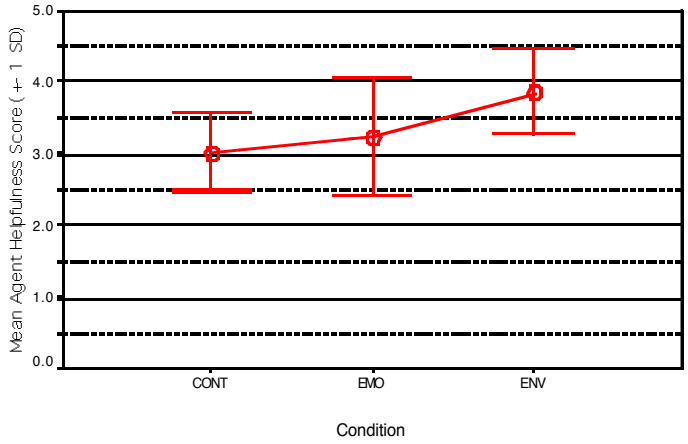


FIGURE 10-5. Mean score for "Agent Helpfulness" (question 3, Evaluation Questionnaire, page 216).

tions were: language understanding (means on a scale from 0 to 10: ENV=7.25, CE=6.67), language use (ENV=7.91, CE=6.83), smoothness of interaction (ENV=6.25, CE=5.41), smoothness of interaction compared to interacting with a dog (means on a scale from 1 to 5: ENV=4.08, CE=3.75), life-likeness compared to any computer character (ENV=3.83, CE=3.16). Comparison of the characters' lifelikeness to a fish in a fishbowl showed a ceiling effect (ENV=4.91, CE=4.83)

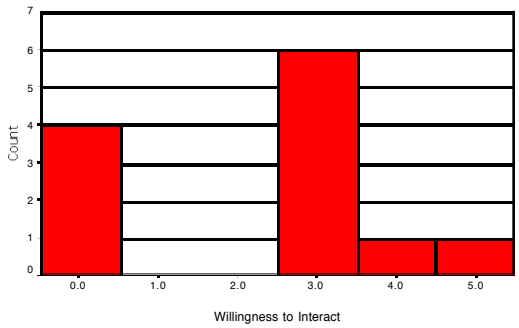


FIGURE 10-6. After interacting with the three characters, the majority of subjects reported that they were more willing to interact with a computer controlled character than before (0 = "No prior opinion", 1 ="Much less willing", 3 = "Equally willing", 5 = "Much more willing", Count = number of subjects).

and was not significant between the conditions, as were none of the other sub-questions. These are summarized in Table 10-3.

Descriptive Statistics

Subjects' reports on three additional variables are worth mentioning: Their answers to the question about increased or decreased willingness to interact with computer controlled characters based on this experience (Prior Beliefs Questionnaire, question 1-b) showed that none of the subjects were less willing, and about half of those with prior ideas about the issue were *more* willing than before (Figure 10-6). The subjects' changed perception of whether machines will ever become intelligent

TABLE 10-3. Items in Question 1, Evaluation Questionnaire for the ENV and CONT+EMO conditions that were *significantly different*. *Scale from 0 to 10; #scale from 1 to 5.

	ENV	CONT+EMO/2
Language Understanding*	7.25	6.67
Language Use*	7.91	6.83
Smoothness of Interaction#	6.25	5.41
Smoothness of Interaction Compared to Interacting with a Dog#	4.08	3.16
Life-likeness Compared to any Computer Character#	3.83	3.16

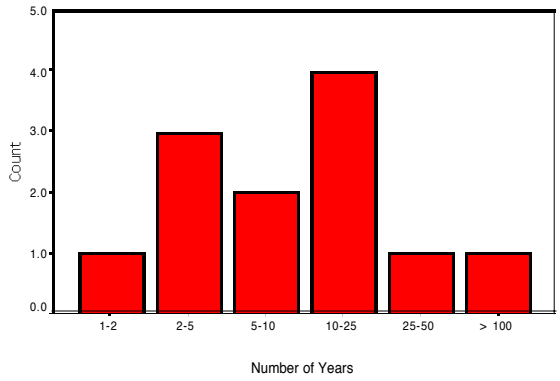


FIGURE 10-8. Distribution of answers for the question "How many years do you think it will take to create a computer controlled character that works perfectly?"

(Prior Beliefs Questionnaire, question 1-d) showed similar trend, with about one-third claiming *increased* confidence in intelligent machines (Figure 10-7).

Their estimation on how long it would take a research team to create a character that "works perfectly" (Prior Beliefs Questionnaire, question 3) showed very positive numbers, the mode being 10-25 years (Figure 10-8), meaning that most subjects, based on this experiment, expect to see characters that "work perfectly" well within their own life-times.

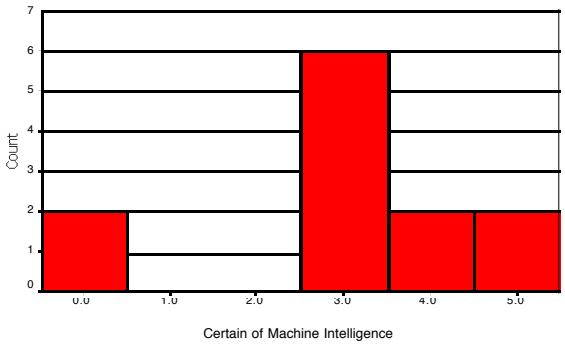


FIGURE 10-7. After interacting with the three characters, no subjects claimed the interaction to have changed their minds about whether machines will ever become intelligent (0 = "No prior opinion", 1 ="Much less certain", 3 = "Equally certain", 5 = "Much more certain", Count = number of subjects).

Answers to Open Questions

Subject responses to the open questions on the questionnaires were illuminating, and are summarized in Figure 10-9.

10.2.5 Discussion

All but two hypotheses where confirmed. This supports the general premise set forth in this experiment—that envelope feedback is important for language based, co-temporal, co-spatial interaction. The two hypotheses that were not confirmed showed a reverse pattern of what was expected, subjects tend to be more hesitant and frustrated in the ENV condition than the other two. A first attempt to answer why this could be might sound like this: Because the ENV condition provides more feedback about the state of the agent's processing, subjects tend to hesitate before speaking, simply because the agent displays behavior that allows them to hesitate in order to minimize overlapping speech. Unfortunately, if this were true, the overlaps in speech should have been the reverse of what they were, i.e. there should be fewer overlaps in the ENV condition than the other two. This was not the case. A more believable explanation to both these reversals is that since the agent's behavior in the ENV condition is more similar to human face-to-face interaction, subjects fell more easily back on a natural interaction style, a more complex one than they exhibited in the other two conditions. And since the characters' perception of the users' actions are limited, it couldn't respond to subtle features in the users' behavior, resulting in more overlaps and hesitations than in the other two conditions. If this is

FIGURE 10-9. Examples of responses to open questions in the Prior Beliefs Questionnaire (Appendix A3.4, page 219).

"Confirmed the idea that computer agents in the future will be able to aid, assist, educate and entertain in everyday life."

"I'm not sure I expect human faces."

"I was hoping that interaction would be faster, fewer pauses between inquiry and response."

"I'm not sure I expect human faces."

"Confirmed the idea that computer agents in the future will be able to aid, assist, educate and entertain in everyday life."

"I thought it would be much colder."

"Roland seemed totally stoned." (Subject in CONT condition)

the case, it points to a need for more sophisticated perceptual mechanisms to support natural, unhindered turn taking and information exchange. Needless to say, Ymir is designed to support such extensions.

Another source of observational evidence supports the above hypothesis. Although the biggest factor by far in determining how much nonverbal behavior the subjects exhibited was personal differences, subjects in the ENV condition tended to look more back and forth between the big screen and the character, tended to gesture more and seemed to be more drawn into the interaction in general. In general, participants tended to mimic the agents' behaviors: If the agent was rigid, they tended to stand still; if the agent was more animated, they tended to be animated. While useful information for future improvement, in the current prototype this could be expected to lead to a less predictable response pattern from the agents, resulting in more errors in judgement of the dialogue state, both of the subject and of the agent.

10.3 Ymir as a Foundation for Humanoid Agent research: Some Observations

One of the main goals in developing the Ymir architecture was to make it suitable as a platform for continued research in humanoid agents. The question then arises, how easy/hard is it to develop new modules, add functionality and modify existing structures? We will look at these questions in turn.

10.3.1 Developing New Modules with the Multimodal Recorder

To deal with the vexing complexity of developing new modules, a multimodal recorder facility was designed (Figure 10-10, Figure 10-13). The Multimodal Recorder allows an agent designer to graphically sketch out any of the internal module's states over time, for any particular period, and to compare events across layers and blackboards. An example of the entire repertoire of Functional Sketchboard[8] messages is shown in Figure 10-13. A real-time display of module states can be viewd in the Module Viewer window (Figure 10-14, page 180). This is very useful for initial testing of modules, to see if they respond correctly to events.

8. The Functional Sketchboard is a blackboard used in the Ymir architecture. It is discussed in Section 7.2.2, page 96.

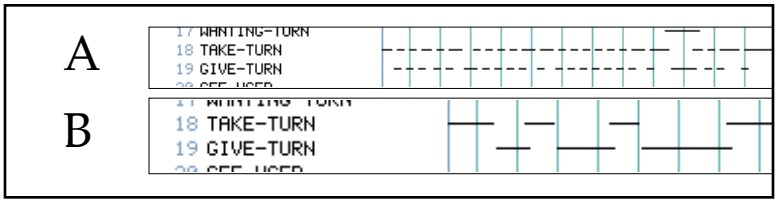
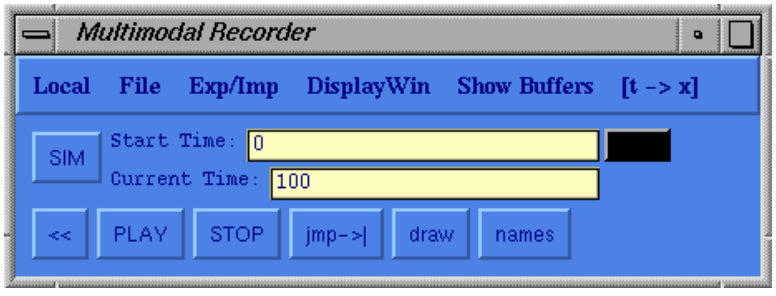


FIGURE 10-11. Using the visualization option of the multimodal recorder, an oscillation problem in the agent's decision pattern for giving and taking the turn (A) was fixed in a matter of hours (B).

Examples of Use

By using this recorder, a feedback problem in the turn taking rules was quickly resolved. The turn taking decision modules of Gandalf were showing a feedback problem causing oscillation in the State Decision Modules between the agent taking turn and giving turn (A, Figure 10-11). The same data from a human subject was fed back using the "simulation" option of the Recorder ("SIM" button in Figure 10-10). This option feeds back data recorded from the sensor and descriptor modules to the decision modules in real-time in the same manner they were originally generated by the human user's actions. The trigger rules for the turn taking modules were modified until a better[9] pattern was achieved (B, Figure 10-11). This modification took less than two hours using the Recorder; the problem had been causing inappropriate behaviors for days before.

FIGURE 10-10. The control panel of the Multimodal Recorder. Menus and buttons allow an agent designer to display the events of an interaction in a graphical format.



9. Notice that there is no "correct" pattern to be achieved here; simply a pattern that allows the agent to respond appropriately.

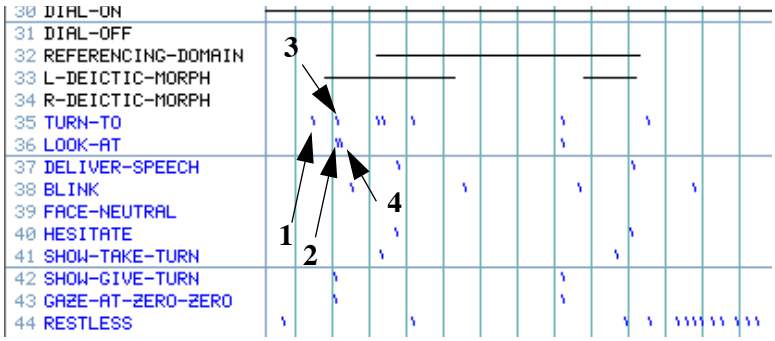


FIGURE 10-12. Example of the morphology (spatial) sensors for deictic gestures turning on. The multimodal descriptor REFERENCING-DOMAIN has more than just the two morphology sensors as input, so it stays on even though the sensor turns off. Gandalf first turns to the user [1], but upon the deictic detection it first looks in the direction of the gesture [2] and then turns in the same direction [3], and immediately readjusts the gaze to fall on the object [4].

Somewhat more difficult was the task of adding a PCL decision module that produces a "problem report" when speech is not recognized. The module (Table 9-4 on page 138, module 4) had to be highly constrained in its trigger and re-trigger conditions to work properly. It took about 5 days to get the right combination of conditions. Without the recorder, or some similar visualization tool, the construction of this module might have been all but impossible.

10.3.2 Adding Functionality: Deictic Gesture at the Input

We will now take an example of the task of adding the perception of, and ability to respond to, deictic manual gestures.

We begin by adding virtual sensors for the simple morphology of outstretched arm and hand above waistline. We come up with two sensors (Table 10-4, 1 & 2), one for each arm (this duplication of arms is a result of the sensing hardware used—other gadgets such as cameras might need a different breakup at this level). Because of the simulated parallel implementation of Ymir, little consideration needs to be paid to special scheduling of the various modules and processes when designing a humanoid, which leaves the implementer free to focus on other issues. Two procedures are called to compute the necessary values, one relating hand position to the trunk, the other comparing the elbow angle to a threshold. If we wanted to get detailed we could add an extended

index finger as yet another hint of deictic morphology. This data is used by a single Multimodal Descriptor (Table 10-4, 3) to determine if the user is addressing the work space.

Adding two new Decision Modules (Table 10-4, 4 & 5), the modules' Expected Lifetime can now be used to decide whether to respond to a deictic gesture, once detected, or to ignore it because it was detected to late.

Figure 10-12 shows part of the pattern created during a long deictic gesture and a shorter one. The speed of the whole loop, from perception to action, is the determinant of whether the character responds at all to the gesture.

10.3.3 Summary

Designing the first character in Ymir took somewhere between one and two months. This number may be expected to go down for second and third character, since much of the modules of the first one stay the same. The prototyping was made possible by using a multimodal recorder and visualization device which allowed the designer to graph internal events over time, at multiple resolutions. To achieve consistency in the internal workings of a character, a designer needs to be careful about stick-

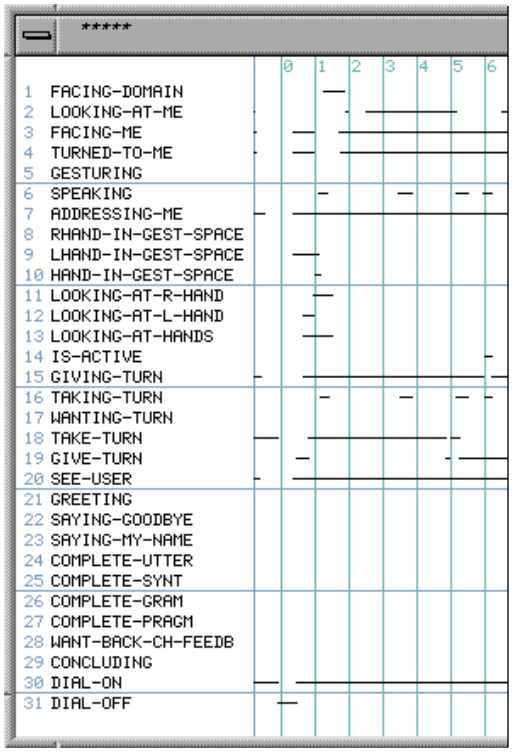


FIGURE 10-13. Example of the Multimodal Recorder display. It shows states of modules in the Reactive Layer over a period of 6 seconds. Lines mean that modules are true; where nothing is drawn they are false. Menu selections allow the user to switch between various sets of modules. In this example, the rules for Gandalf were being modified to produce the correct states. (See Tables 9-1, 9-3 & 9-3 in Chapter 9.)

NAME: l-deictic-morph TYPE: body-sensor-var-ref DATA-1: nil DATA-2: nil INDEX-1: (**Get-Body-Part** Left-Arm-Index) INDEX-2: (**Get-Trunk-Dir**) FUNC: **Deictic-Sketch**	**1**
NAME: r-deictic-morph TYPE: body-sensor-var-ref DATA-1: nil DATA-2: nil INDEX-1: (**Get-Body-Part** Right-Arm-Index) INDEX-2: (**Get-Trunk-Dir**) FUNC: **Deictic-Sketch**	**2**
NAME: referencing-domain POS-CONDS: (l-deictic-morph 0.7)(r-deictic-morph 0.8)(facing-domain 0.5)(speaking 0.5)(facing-me 0.4) NEG-CONDS: nil THRESH: 1.0	**3**
NAME: look-where-pointing TYPE: RL-Ext-Dec-Mod EL: 200 MSGS: (look-at 'big-screen) POS-CONDS: (referencing-domain) NEG-CONDS: nil POS-RESTR-CONDS: nil NEG-RESTR-CONDS: (referencing-domain)	**4**
NAME: turn-to-where-pointing TYPE: RL-Ext-Dec-Mod EL: 200 MSGS: (turn-to 'big-screen) POS-CONDS: (referencing-domain)(**FS-time-since** 'referencing-domain 100) NEG-CONDS: nil POS-RESTR-CONDS: nil NEG-RESTR-CONDS: (referencing-domain)	**5**

TABLE 10-4. The function **Deictic-Sketch** uses morphology to find deictic gestures. It checks the angle of the elbow and the height of the hand above the waistline to determine if the posture of an arm might be doing a deictic gesture. Additional information such as posture of the hand could increase the accuracy of this virtual sensor.

ing to the architecture of Ymir. When this is done, however, Ymir provides a powerful foundation for designing and expanding the design, of communicative characters.

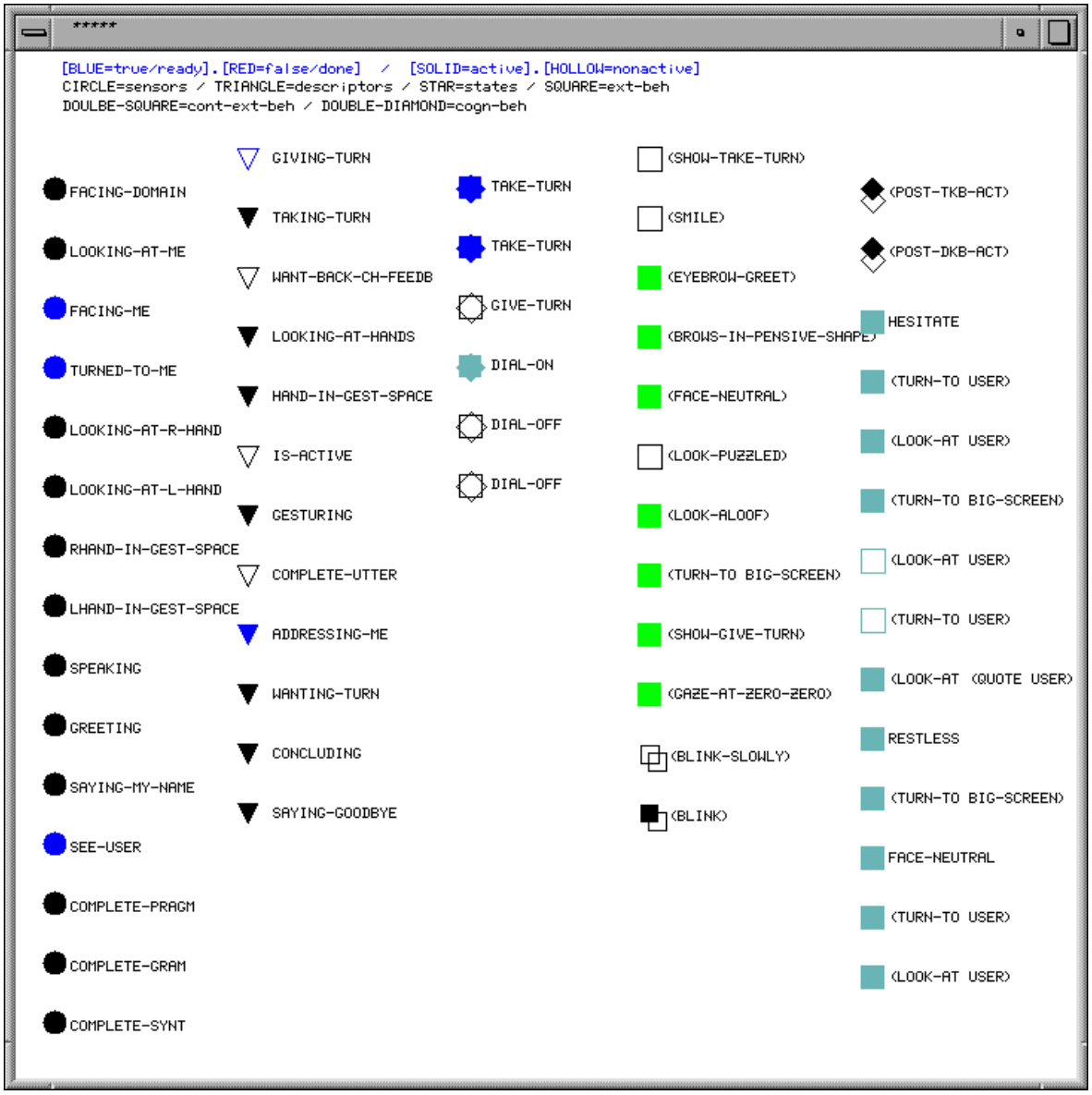


FIGURE 10-14. The Module Viewer allows a character designer to view the states of the modules designed, color-coded and updated in real-time.