AI Magazine Volume 27 Number 2 (2006) (© AAAI)

Toward Virtual Humans

William Swartout, Jonathan Gratch, Randall W. Hill, Eduard Hovy, Stacy Marsella, Jeff Rickel, and David Traum

■ This article describes the virtual humans developed as part of the Mission Rehearsal Exercise project, a virtual reality–based training system. This project is an ambitious exercise in integration, both in the sense of integrating technology with entertainment industry content, but also in that we have joined a number of component technologies that have not been integrated before. This integration has not only raised new research issues, but it has also suggested some new approaches to difficult problems. We describe the key capabilities of the virtual humans, including task representation and reasoning, natural language dialogue, and emotion reasoning, and show how these capabilities are integrated to provide more human-level intelligence than would otherwise be possible.

Achieving human-level intelligence in cognitive systems requires a number of core capabilities, including planning, belief representation, communication ability, emotional reasoning, and most importantly, a way to integrate these capabilities. And yet, for many researchers, software integration is often regarded as a kind of necessary evil—something to make sure that all the research components of a large system fit together and interoperate properly—but not something that is likely to contribute new research insights or suggest new solutions. We have found, on the contrary, that the conventional wisdom about integration does not hold: as we describe in this article,[1] the integration process has raised new research issues and at the same time has suggested new approaches to long-standing issues. We begin with a brief description of the background behind our work in training and the approach we have taken to improving training. We then describe the technology components we have developed, the system architecture we use, and we conclude with some of the insights we have gained from the integration process.

Virtual Humans for Training

We have been constructing virtual humans to explore research issues in achieving cognitive systems with human-level performance. These issues, which we describe in detail below, span a number of technical areas in artificial intelligence including speech recognition, natural language understanding and generation, dialogue modeling, nonverbal communication, task modeling, social reasoning, and emotion modeling.

Virtual humans are software artifacts that look like, act like, and interact with humans but exist in virtual environments. We have been exploring the use of virtual humans to create social training environments, environments where a learner can explore stressful social situations in the safety of a virtual world.

  0738-4602-2006 / $2.00

Figure 1. The Mission Rehearsal Exercise System.

From left to right: The platoon sergeant, the mother with her injured boy, a medic, and a crowd.

We designed the Mission Rehearsal Exercise (MRE) system to demonstrate the use of virtual human technology to teach leadership skills in high–stakes social situations. MRE places the trainee in an environment populated with virtual humans. The training scenario we are currently using is situated in a small town in Bosnia. It opens with a lieutenant (the trainee) in his Humvee. Over the radio, he gets orders to proceed to a rendezvous point to meet up with his soldiers to plan a mission to assist in quelling a civil disturbance. When he arrives at the rendezvous point, he discovers a surprise (see figure 1). One of his platoon's Humvees has been involved in an accident with a civilian car. There is a small boy on the ground with serious injuries, a frantic mother, and a crowd is starting to form. A TV camera crew shows up and starts taping. What should the lieutenant do? Should he stop and render aid? Or should he continue on with his mission? Depending on decisions he makes, different outcomes will occur.

Our virtual humans build on prior work in the areas of embodied conversational agents (Cassell et al. 2000) and animated pedagogical agents (Johnson, Rickel, and Lester 2000), but they integrate a broader set of capabilities than any prior work. For the types of training scenarios we are targeting, the virtual humans must integrate three broad influences on their behavior: they must perceive and act in a three-dimensional virtual world, they must engage in face-to-face spoken dialogues with people and other virtual humans in such worlds, and they must exhibit humanlike emotions. Classic work on virtual humans in the computer graphics community focused on perception and action in three-dimensional worlds (Badler, Phillips, and Webber 1993; Thalmann 1993), but largely ignored dialogue and emotions. Several systems have carefully modeled the interplay between speech and nonverbal behavior in face-to-face dialogue (Cassell et al. 2000; Pelachaud, Badler, and Steedman 1996) but these virtual humans did not include emotions and could not participate in physical tasks in three-dimensional worlds. Some work has begun to explore the integration of conversational capabilities with emotions (Lester et al. 2000; Marsella, Johnson, and LaBore 2000; Poggi and Pelachaud 2000), but still does not address physical tasks in three-dimensional worlds. Likewise, prior work on Steve addressed the issues of integrating face-to-face dialogue with collaboration on physical tasks in a three-dimensional virtual world (Rickel and Johnson 2000), but Steve did not include emotions and had far less sophisticated dialogue capabilities than our current virtual humans. The tight integration of all these capabilities is one of the most novel aspects of our current work.

The virtual humans, including the sergeant, medic, and mother in the scenario described in the previous section build on the earlier Steve system. Their behavior is not scripted; rather, it is driven by a set of general, domain-independent capabilities discussed below. The virtual humans perceive events in the simulation, reason about the tasks they are performing, respond verbally through generated speech and nonverbally through gestures and facial expressions, and react emotionally to events as they unfold.

Integration Issues

In order for virtual humans to collaborate with people and each other in scenarios like the peacekeeping mission, they must include a wide variety of capabilities, such as perception, planning, spoken dialogue, and emotions. Creating an integrated virtual human that could support such a broad range of behaviors presented some significant challenges both for the integrating architecture and the software development process.

One challenge was that each of the major components (such as dialogue management or emotion modeling) was developed by a different research team. Since the MRE effort in total involved dozens of people, we felt it was necessary to break things down in this way to keep each task manageable. A second challenge was that each research team was attempting to advance the state of the art for their component—to do things that had not been done before. Thus, capabilities that had not been available when the research started might become available as it progressed.

Taken together, these challenges meant that it was not possible to determine in advance what information one module might be able to provide to another, or even what information would be needed. This meant that a top-down design was not possible; instead the design emerged as the research progressed. Another consequence of using separate teams was that bugs arising from interdependencies between modules were often difficult to track down.

Another challenge came from the real-time nature of the interactive application. Because the agents were reacting to events as they unfolded, small timing differences between two runs could result in different behaviors, even when the same inputs were used for each run. These behavioral differences sometimes made it difficult to duplicate and debug problems with the agents.

In the next two sections, we describe the virtual human architecture and the software development process we used to help ameliorate these issues. That is followed by an overview of the components in the virtual human system. Finally we conclude with a discussion of the synergies that have emerged from integration and some of the lessons we have learned about system integration on this scale.

Architecture

As we argued in the previous section, a top-down design was not possible, and it was not possible to determine in advance all the inputs and outputs for the various components. To provide the needed flexibility, we used a blackboard architecture, in which memory is shared and individual components have access to the intermediate and final results of other components by default. The alternative, in which each module would explicitly pass specific information to other components, would require constant revision as we made progress understanding the interdependencies among components.

For our integrated architecture, we chose Soar (Newell 1990), because it allows each component to be implemented with production rules that read from and write to a common working memory, which acts as the desired blackboard. Soar further breaks computation into a sequence of intermediate *operators* that are proposed in parallel but selected sequentially through an arbitration mechanism. This allows for tight interleaving of operators from individual components and flexible control over their priority. They use the communications bus (see figure 2) to send messages to one another, to the character bodies, and to the audio system.

All components of the virtual humans are implemented in Soar, with several exceptions: speech recognition, natural language understanding (syntactic and semantic analysis), synchronization of verbal and nonverbal components of output utterances, and speech synthesis. It was less practical to implement these four components in Soar because each was built on top of existing software that would have been difficult to reimplement. In addition, these modules also work roughly as pipelines, with well-defined inputs and outputs, so the flexibility that Soar provides was less necessary for these components.

Virtual Human Components

Task Representation and Reasoning. To collaborate with humans and other synthetic teammates, virtual humans need to understand how past events, present circumstances, and future possibilities impact team tasks and goals. For example, the platoon sergeant agent in figure 1 must be able to brief the trainee on past events that led to the accident as well as how the victim's current injuries impact the platoon's future mission. More generally, agents must understand task goals and how to assess whether they are currently satisfied, the actions that can achieve them, how the team must co-

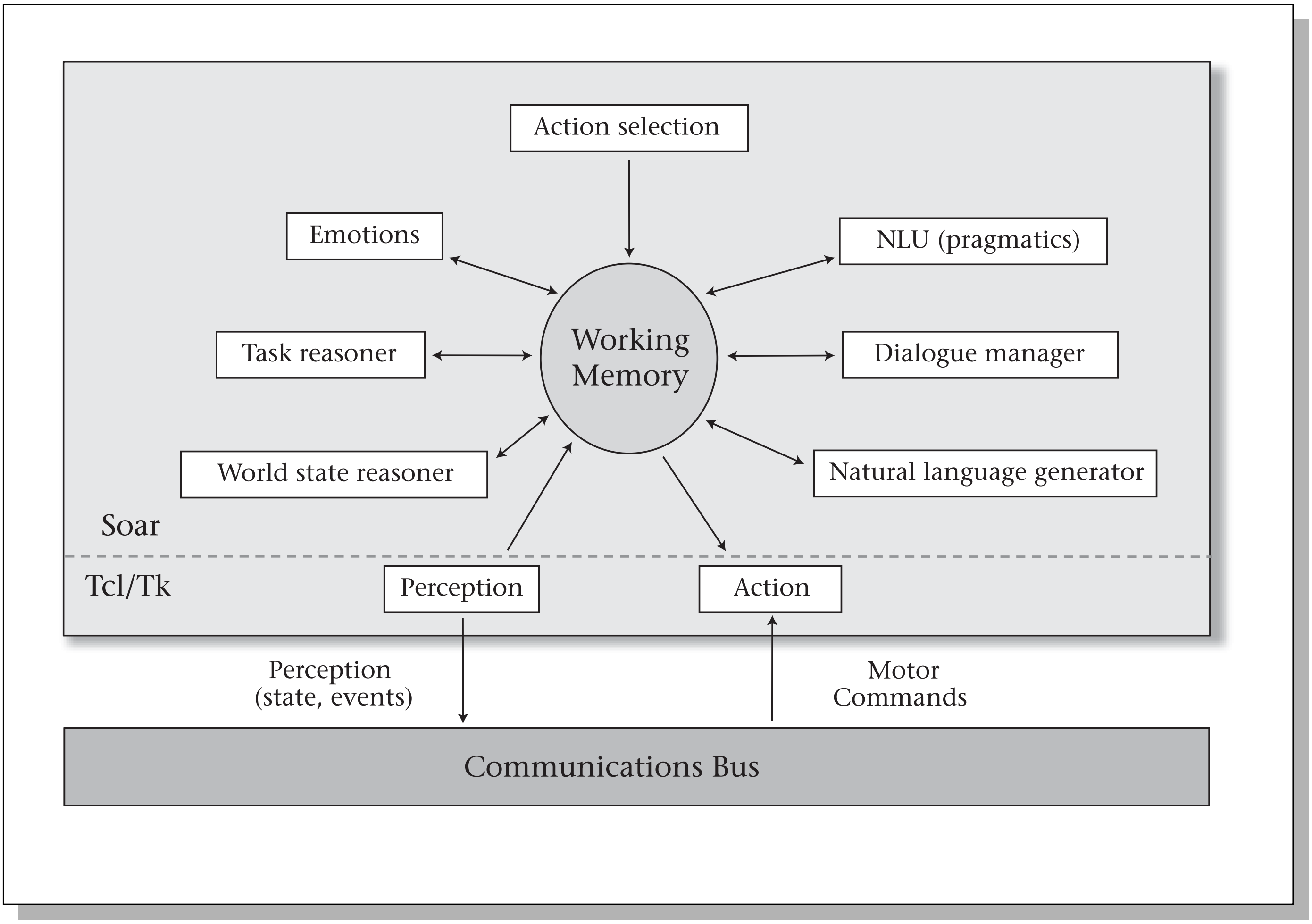

Figure 2. Virtual Human Architecture.

ordinate the selection and execution of those actions, and how to adapt execution to unexpected events.

To provide this understanding, we employed a model-based programming approach and constructed an explicit model of the tasks, events, and goals in the domain. Because this model can be shared among system components, it directly addresses some of the issues that arise in coordination and knowledge sharing when the components of the virtual human are integrated.

Agents use domain-independent reasoning algorithms operating over a domain-specific declarative representation of team tasks. The representation incorporates elements of decision-theoretic plan representations (allowing agents to reason about the utility and likelihood of future possibilities) with an explicit representation of beliefs and intentions (important for multiagent reasoning). This representation is divided into explicit representations of past episodes, present state, and future task-related information: The *causal history* maintains a sequence of past observed steps (including unexpected and nontask events) and interdependencies between past steps and present or future states (such as causal links). The *current world description* represents the current state of the world through a list of propositions. The *task description* includes of a set of possible future steps, each of which is either a primitive action (for example, a physical or sensing action in the virtual world) or an abstract action which must itself be further decomposed. Abstract actions give tasks a hierarchical structure. Interdependencies are represented as a set of ordering constraints, causal links and threat relations.

In addition to understanding the structure of tasks, agents must understand the roles of each team member. Each task step is associated with the team member that is responsible for performing it as well as a possibly different agent that has authority over its execution; that is, the teammate responsible for a task step cannot perform it until authorization is given by the specified teammate with authority (Traum et al. 2003). This is required to model the hierarchical organizational structure of some teams, such as in the military.

An agent's task model represents its understanding of the task in general, independent of the current scenario conditions (different agents may have different representations of the same task). Agents continually monitor the state of the virtual world through messages from the simulator that are filtered to reflect perceptual limitations (Rickel et al. 2002) and update their plans accordingly. The result of this planning algorithm specifies how the agent privately believes that the team can collectively complete the task, with some causal links specifying the interdependencies among team members' actions.

A key aspect of collaborative planning is negotiating about alternative ways to achieve team goals (Traum et al. 2003). To support such negotiation, the decision-theoretic planner can reason about alternative, mutually exclusive courses of action (recipes) for achieving tasks, their likelihood, and the utility of certain consequences, allowing the system to assess the relative strengths and weaknesses of different alternatives. These courses of action are self-contained hierarchical tasks in the sense defined above, and subject to the same dynamic task reasoning. For example, one might evacuate someone to a hospital by using either a medevac helicopter or an ambulance. Depending on the circumstances, only one option might be possible (for example, the medevac may be unavailable or the injuries may be too severe for an ambulance), but if both are valid options, they must be ranked through some reasoned analysis of their relative costs and benefits.

Natural Language Dialogue. In many ways, our natural language processing components and architecture mirror fairly traditional dialogue systems. There is a speech recognizer, semantic parser, dialogue manager, NL generator, and speech synthesizer. However, the challenges of the MRE project, including integration within an immersive story environment as well as with the other virtual human components required innovations in most areas. Here we briefly describe the natural language processing components and capabilities; we will return later to some of the specific innovations motivated by this integration.

The speech recognizer was built using Sonic (Pellom 2001), with a domain specific n-gram language model and with locally trained acoustic models (Wang and Narayanan 2002). Output is currently the single best interpretation, as well as indications of when the user starts and stops speaking, to manage gaze control and turn-taking behavior of agents.

Speech recognition output is processed by the semantic parser module, which produces a semantic representation of the utterances. The parser combines two finite-state transducers and a statistically trained processing engine, each of which produces candidate semantic interpretations for the incoming word stream, from which a best-guess is then selected (Feng and Hovy 2003). In cases in which perfect and standard or expected input is received, the finite state transducers provide very accurate output; when imperfect input is given, the statistical engine will robustly produce representations that may possibly be incomplete or partially incorrect. The module will provide addressee information (if vocatives were present), sentence mood, and semantic information corresponding to states and actions related to the task model (Traum 2003).

The output of the speech recognizer and semantic parser is passed to the Soar-based dialogue management system for each virtual human agent. This information is then matched against the agent's internal representation of the context, including the actions and states in the task model, current expectations, and focus to determine a set of candidate interpretations. These interpretations may be underspecified, due to impoverished input, or overspecified in cases of incorrect input (either an out-of-domain utterance by the user, or an error in the speech recognizer or semantic parser). In some cases, underspecified elements can be filled in with reference to the agent's knowledge; if not, the representation is left underspecified and processing continues. Each agent's dialogue component also produces a set of dialogue act interpretations of the utterance. Some of these are traditional speech acts (for example, assert, request, info-request) with content being the semantic interpretation, while others represent other levels of action that have been performed, such as turn-taking, grounding, and negotiation (Traum and Rickel 2002).

Dialogue management follows the approach of the TRINDI Project (Larsson and Traum 2000), and specifically the EDIS system (Matheson, Poesio, and Traum 2000). Dialogue acts are used to update an *Information State* that is also used as context for other aspects of agent reasoning (Traum and Rickel 2002). Decisions of how to act in dialogue are tightly coupled with other action selection decisions in the agent. The agent can choose to speak, choose to listen, choose to act related to a task, and so

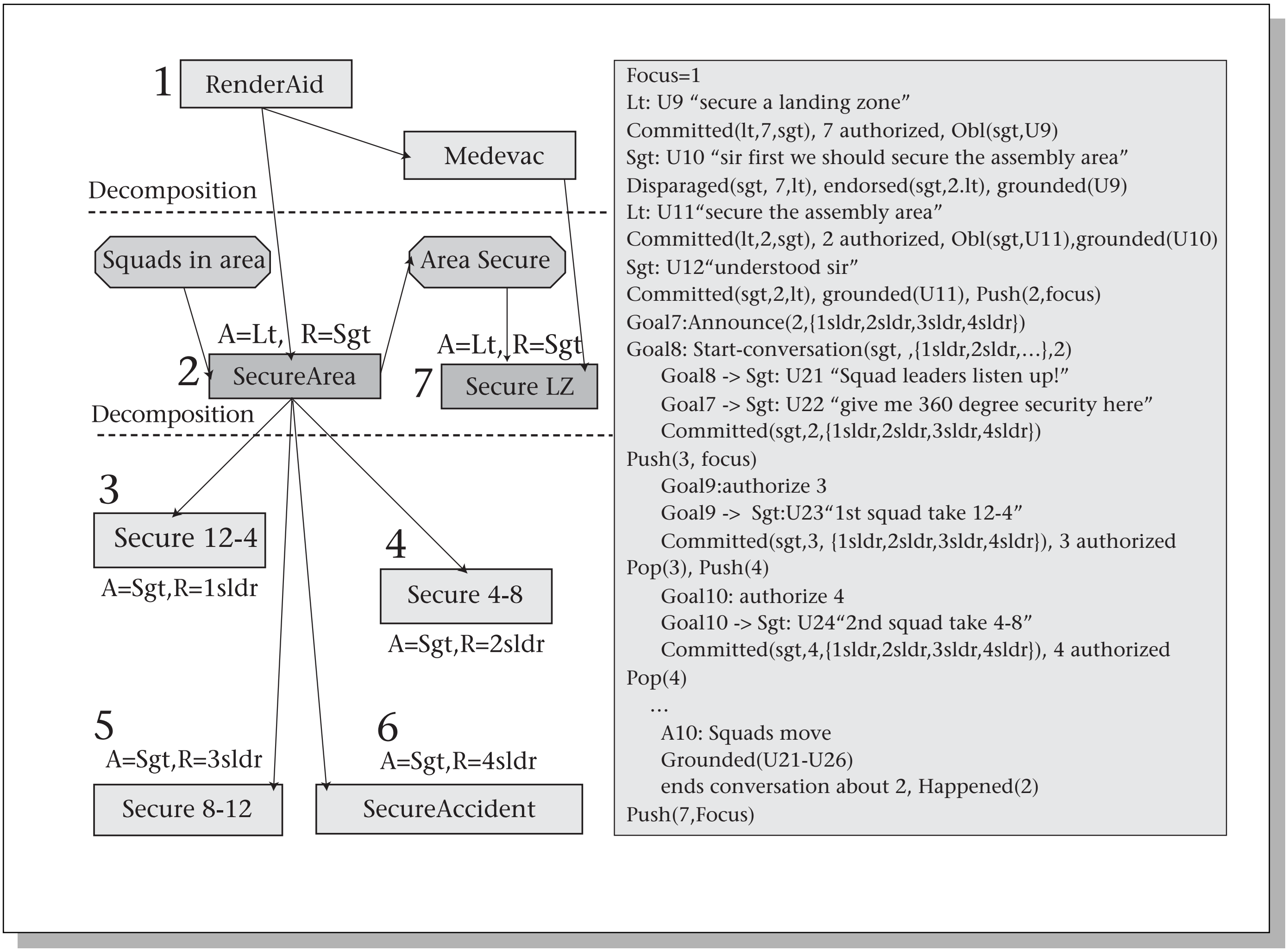

Figure 3. Sample Task Model and Dialogue Interaction.

on. Aspects of the information state provide motivations to speak, including answering questions, negotiating with respect to a request or order, giving feedback of understanding (acknowledgements, repairs, and repair requests), and making suggestions and issuing orders, when appropriate according to the task model.

Once a decision is made to speak, there are several phases involved in the language production process, including *content selection, sentence planning*, and *realization*. The final sentence is then augmented with communicative gestures and sent to the synthesizer and rendering modules to produce the speech. Meanwhile, messages are sent to other agents, letting them know what the agent is saying (Fleischman and Hovy 2002). The speech synthesizer uses Festival and Festvox, with locally developed unit-selection limited-domain voices to provide the emotional expressiveness needed to maintain immersiveness (Johnson et al. 2002).

Figure 3 shows a brief example of how dialogue behavior is integrated with task reasoning. The left side of the figure shows a small fragment of the task model: part of the "Render aid" task involves securing the assembly area, which requires that the squads are in the area; it has a decomposition involving actions of various squads, and has the effect that the area is secure. The figure also shows which agents are responsible (R) for seeing that an action is performed (either doing it themselves or acting as team leader making sure the subtasks are carried out), and which agents have authority (A) to have the action performed. With reference to this piece of the task model, consider the dialogue fragment on the right. Initially the focus is on the render aid task. When the lieutenant issues the command to secure the area (utterance U11), the sergeant recognizes the command as referring to a subaction of Render Aid in the current task model (Task 2). As a direct effect of the lieutenant issuing a command to perform this task, the lieutenant has committed himself to the task, the sergeant has an obligation to perform the task, and the task becomes authorized. Because the sergeant already agrees that this is an appropriate next step, he is able to accept it with utterance U12, which also commits him to perform the action. The sergeant then pushes this task into his task model focus and begins execution. In this case, because it is a team task requiring actions of other teammates, the sergeant, as team leader, must announce the task to the other team members. Thus, the system forms a communicative goal to make this announcement. Before the sergeant can issue this announcement, he must make sure he has the squad leaders' attention and has them engaged in conversation. He forms a goal to open a new conversation so that he can produce the announcement. Then his focus can turn to the individual tasks for each squad leader. As each one enters the sergeant's focus, he issues the command that commits the sergeant and authorizes the troops to carry it out. When the sergeant observes the troops move into action, he can infer that they have understood his order and adopted his plan. When the task completes, the conversation between sergeant and squad leaders finishes and the sergeant turns his attention to other matters.

Emotion. As our agents attempt to realistically model the behavior of humans in high-stress scenarios, it is important to model the role emotion plays in influencing decision making and behavior. Our work on modeling emotion is motivated by *appraisal theory*, a psychological theory of emotion that emphasizes the relationship between emotion and cognition (Lazarus 1991). The theory posits two basic processes: *Appraisal* generates emotion by assessing the person-environment relationship (did an event facilitate or inhibit the agent's goals; who deserves blame or credit). *Coping* is the process of dealing with emotion, either by acting externally on the world (problem-focused coping), or by acting internally to change beliefs or attention (emotion-focused coping). Coping and appraisal interact and unfold over time, modeling the temporal character of emotion noted by several emotion researchers (Lazarus 1991; Scherer 1984): an agent may "feel" distress for an event (appraisal), which motivates the shifting of blame (coping), which leads to anger (reappraisal).

In recasting this theory as a computational model, we exploit the agent's automated reasoning capabilities as proxies for the cognitive mechanisms that underlie emotion (Gratch and Marsella 2004). For example, to distinguish joy from distress, the agent must assess the valence of an event, something supported by the agent's decision-theoretic reasoning; to distinguish distress from anger, it must assess whether a threatening act by another was foreseen and intentional, something supported by the agent's ability to reason about beliefs and intentions; to assess the potential to cope with an emotional event, the agent must be able to reason about its ability to plan or seek support from others, something supported by the task and dialogue reasoning.

Our approach to appraisal assesses the agent-environment relationship through features of this explicit task representation (Gratch 2000). Speaking loosely, we treat appraisal as a set of feature detectors that map features of the task and dialogue state into appraisal variables that characterize the consequences of an event from the agent's perspective. These variables include the desirability of those consequences, the likelihood of them occurring, who deserves credit or blame, and a measure of the agent's ability to alter those consequences. The result is one or more *appraisal frames* that characterize the agent's emotional reactions to an event.

Our computational model of coping (Marsella and Gratch 2002) similarly exploits the task and dialogue representations to uncover which features led to the appraised emotion, and what potential there may be for altering these features. In essence, coping is the inverse of appraisal. Coping operates on the same representations as the appraisals, but while appraisal looks at changes in the world and beliefs to determine their effect on emotion, coping seeks to reduce (primarily negative) emotions by making changes to the world or beliefs. There are two broad classes of coping strategies. One works by making changes in the world. For example, a person might be driving and see an accident. Feeling upset, he could cope with his emotion by making a cell phone call to get help. The other broad class of coping strategies operates not on the world but on internal beliefs. Using the same example, rather than placing a cell call, the driver could decide that the accident victim was careless and got what he deserved. The driver would still feel better, but his coping strategy would affect only his own beliefs. Our coping strategies can involve a combination of such approaches. This mirrors how coping processes are understood to operate in human behavior whereby people may employ a mix of problem-focused coping and emotion-focused coping to deal with stress.

Action and Body Movements. Internally, the virtual humans are perceiving events, understanding utterances, updating their beliefs, formulating and revising plans, generating emotional appraisals, and choosing actions. Agents manifest the rich dynamics of their cognitive and emotional inner state through external behavior using the same verbal and nonverbal cues that people use to understand one another, and these behaviors must be seamlessly integrated across modality and across time.

Here we summarize the model discussed in Marsella, Gratch, and Rickel (2003), which drives gaze, facial expressions, and body gestures based on features of the agent's dynamic cognitive state. Gaze indicates a character's focus of attention and is synchronized to the character's inner thoughts. For example, task-related behaviors (such as monitoring for an expected effect or action) trigger a corresponding gaze shift, and gaze during social interactions is driven by the dialogue state and the state of the virtual human's own processing (for example, gaze at an interlocutor who is speaking, gaze aversion during utterance planning to hold the turn). Facial expressions both convey emotion and augment verbal communication. In humans, these behaviors can be used intentionally by an individual to inform or deceive but can also unintentionally reveal information about the individual's mental state, and our work integrates these aspects: by tying some expressive behavior to emotional appraisal we reveal "true" mental state, whereas tying other behaviors to coping strategies, we inform intentional displays. Finally, a wide range of body movements emphasize and augment speech. Our approach plans the utterance, annotates it with nonverbal behavior, then passes it to a text-to-speech system that schedules both the verbal and nonverbal behavior, using BEAT (Cassell, Vilhjálmsson, and Bickmore 2001), although we augment this to express not only the syntactic, semantic, and pragmatic structure of the utterance, but also emotional appraisal and coping information as well.

Putting It Together: The Value of Integration

We have described the major technical components of the virtual humans. As we pointed out in the introduction, software integration is necessary to make sure that all the various pieces in a system work together properly, but one usually expects that the real research takes place in building the individual components. One does not expect to learn much from integration (except perhaps to find that some components do not interface properly). However, in integrating the Mission Rehearsal Exercise system, we have been surprised: we have uncovered new research issues and some new approaches to existing problems have been suggested. In this section we outline some of the things we learned as we brought all the pieces together.

The Pervasive Effect of Emotion. In humans, emotion has a broad effect on behavior. It affects how we speak, how we gesture, our posture, and even how we reason. And, of course, emotion is indispensable for creating good stories and compelling characters. In integrating emotion into our virtual humans, we have found that we need to deal with a similarly broad range of issues. Models of emotion can both affect the behavior of other components of the virtual human and provide additional knowledge that the system can use in reasoning. Below we give an example of each.

Emotionally Appropriate Natural Language Generation. A big challenge for natural language generation in MRE is the generation of emotionally appropriate language, which expresses both the desired information and the desired emotional attitude towards that information. Each expressive variant casts an emotional shade on each representational item it contains (for example, the phrase governed by the verb "ram" as in "They rammed into us, sir" casts the subject in a negative and the object in a positive light). Prior work on the generation of variation expressions, such as Bateman and Paris (1989) and Hovy (1990), uses quite simplistic emotional models of the speaker and hearer. In general, these systems simply had to choose among a small set of phrases, and within the phrase from a small set of lexical fillers for certain positions of the phrase, where each alternative phrase and lexical item was preannotated with an affective value such as *good* or *bad*.

The presence in MRE of an emotion model provides a considerably finer-grain level of control, enabling principled realization decisions over a far more nuanced set of expressive alternatives. Given many representational items, a rich set of emotional values potentially holding for them, and numerous phrases, each with its own combination of positive and negative fields, the problem was to design a system that can reliably and quickly find the optimal phrasing without dropping content. To compute shades of connotation more accurately and quickly, we created a vector space in which we can represent the desired attitudes of the speaker (as specified by the emotion model) as well as the overall emotional value of each candidate expression (whether noun phrase or whole sentence). Using a standard Euclidean distance measure, we can then determine which variant expression most closely matches the desired effect. See Fleischman and Hovy (2002) for details.

Using Emotion to Determine Linguistic Focus. In natural language, we often refer to things in imprecise ways. To correctly interpret such referents in a natural language utterance, one needs to understand what is in linguistic focus. Loosely speaking, one needs to understand what is the main subject of discussion. For example, when the lieutenant trainee arrives at the accident scene in the MRE scenario, he might ask the sergeant, "What happened here?" In principle many things have happened: the lieutenant just drove up, the soldiers assembled at the meeting point, an accident occurred, a crowd formed, and so forth. The sergeant could talk about any one of these and be factually correct, but not necessarily pragmatically appropriate. A number of heuristics have been developed to model linguistic focus. One such heuristic is based on the idea of recency. It holds that the entity that is in linguistic focus is whatever was most recently discussed, or occurred most recently. In this case, recency does not work, since the sergeant would sound quite silly if he responded: "Well, you just drove up, sir." On the other hand, people are often focused most strongly on the things that upset them emotionally, which suggests an emotion-based heuristic for determining linguistic focus. Because we have modeled the sergeant's emotions in MRE, the dialogue planning modules that have access to the fact that he is upset about the accident can use that information to give the most appropriate answer: describing the accident and how it occurred.

Integration Lessons Learned

As we built the MRE system, we found that in many ways the *process* used in constructing the system could be just as critical as the system's architecture to the success of the endeavor. In this section we summarize some of the lessons we learned along the way.

Integrate early, and often. Because a top-down design was not possible, we found that it was important to begin integration testing early, even before any of the components had reached full functionality. In that way, it was possible to identify unanticipated conflicts and lacunae earlier in the development process, making them easier and less costly to correct. In addition, we found that it was important to continue to perform integration tests on a regular and frequent basis. We performed integration tests roughly every two weeks. Again, the frequent tests allowed us to identify and correct problems early on.

Version control software is essential. While many small research projects can be successfully executed without the need for version control software we found it to be essential due to the fact that many semi-independent teams were integrating their software results together. Without some sort of version control it would have been easy to mix incompatible software modules inadvertently.

It cannot all be research; use existing components where possible. Due to the uncertainty in the design and relative immaturity of research components, each research module adds to the risk of the integrated system. To reduce risk in MRE, we used existing components and frameworks wherever possible and created research components only when the capabilities we needed were not available.

Move from heterogeneous to homogeneous platforms. Early on, because we wanted to make use of existing software to prototype the MRE system rapidly, the MRE system used a broad range of hardware platforms, including an SGI IR3, Macs, and PCs, reflecting which platforms the software had originally been written on. While this allowed us to get a version of the system running rapidly, it introduced reliability issues and also meant that running the system required considerable expertise on the various platforms. We have since moved to a more homogeneous platform, standardizing on PCs, which has increased reliability. We believe that the original decision to adopt a heterogeneous approach was correct, but it was also necessary to make the transition as the system matured.

Component "stand-ins" are needed. We found that between integration tests, it was often difficult for developers to do meaningful tests on their components if the components were running independently. We found it useful to develop stand-in components that mimicked the I/O behavior of real components even through simpler methods. For example, we developed a "fake speech recognizer" that allowed a developer to type in or load previously recognized text, bypassing speech recognition, and we developed a simple rule-based simulator that mimicked the agent's interface with the virtual environment. Such "mimics" could be hooked up to the rest of the system to allow the remaining components to be tested in context. This approach improved productivity between integration tests.

Most important: a shared vision. Building a large-scale integrated system that combines a number of research components is not easy. In managing this process perhaps the most critical thing is that the whole team must have a

Figure 4. Negotiating with a Doctor.

shared vision: they must see the value and expected results of the integration effort. It is that shared vision that will keep the team working together and making progress during difficult times.

MRE Status and Evaluation

An initial version of the MRE system described in this article has been implemented and applied to the peacekeeping training scenario described earlier. The system allows the trainee, playing the role of the lieutenant, to interact freely (through speech) with the three virtual humans (sergeant, medic, and mother). The trainee takes action in the virtual world through commands to the sergeant, who in turn commands the squads. Ultimately, the experience terminates with one of four possible endings, depending on the trainee's actions. However, unlike interactive narrative models based on an explicit branching structure, the system does not force the trainee through a predetermined sequence of decision points, each with a limited set of options; the trainee's interactions with the characters is unconstrained and limited only by the characters' understanding and capabilities.

The understanding and capabilities of the virtual humans is limited by the coverage of their spoken dialogue models and their models of the domain tasks. The sergeant's speech recognizer currently has a vocabulary of a few hundred words, with a grammar allowing recognition of 16,000 distinct utterances. His natural language understanding module can currently produce semantic representation frames for all of these sentences as well as providing (sometimes partial) results for different or ill-formed input. His natural language generation module currently expresses all communicative goals formed by the dialog module, modulating some of them for affective appropriateness. His speech synthesis module currently has a vocabulary of more than 1000 words. The sergeant's domain task knowledge, which is the most complex among all the virtual humans in the scenario, includes about 40 tasks, and about 150 properties of the world. While the tasks represent the full range of actions that the sergeant can understand and carry out, his ability to talk about these tasks and properties (for example, answer questions and give advice) is broad, limited only by the coverage of the spoken dialogue modules as described above.

Despite its complexity, real-time performance of the system is good, although we are continuing to improve latencies. Given an utterance by the user, a virtual human typically responds within 3 seconds, including speech recognition, natural language understanding, updating dialogue and emotional states, choosing how to respond, natural language generation, planning the voice output and accompanying gestures and visemes, and finally producing the speech. As is typical of humans, the virtual humans are producing communicative behaviors throughout this time delay, including averting gaze from the user during the utterance planning phases to indicate that they are formulating a response (Kendon 1967).

We have tested the system with a variety of users acting as trainees, including subjects with and without prior knowledge of the military domain. Not surprisingly, subjects with military knowledge were substantially more successful, since they understood the context and how to proceed. Initial evaluation results and metrics of dialogue interaction using military cadets are presented in (Traum, Robinson, and Stephan 2004)

Negotiation: A New Domain

Recently, we have ported our virtual humans to a new application domain that is intended to teach trainees skills in negotiation. The trainee plays the part of an army captain whose mission is to pursue a medical relief doctor to change the location of his clinic. The medical doctor is played by a virtual human (see figure 4) and has been designed to resist negotiation in the way that psychologists have found that people resist negotiation (see Traum et al. 2005). An initial version of this new system was implemented in about 90 days, and while the new domain naturally required new art assets and new task models, about 80 percent of the general-purpose virtual human code used in MRE was reused in this new application. We feel that the speed of implementation and degree of code reuse provide evidence of the flexibility and robustness provided by our architectural designs.

Human-level intelligence requires a number of core capabilities, including planning, belief representation, communication ability, emotional reasoning, and most importantly, a way to integrate these capabilities. The virtual humans in the MRE project represent a significant step along this path.

Acknowledgements

Jeff Rickel created much of the virtual human architecture and contributed to most aspects of the project. Sadly, he died in July 2003. The authors would like to thank the many people who collaborated on the virtual humans and Mission Rehearsal Exercise project for their hard work and enthusiasm that contributed greatly to this research. The research described here was developed with funding from the U.S. Army Research Development and Engineering Command and funding from the United States Department of the Army under contract number DAAD 19-99-D-0046. Any opinions, findings, and conclusions or recommendations expressed in this article are those of the authors and do not necessarily reflect the views of the United States Department of the Army.

Note

1. This article is based on and expands the discussion of issues in integration presented in Swartout et al. 2005.

References

Badler, N. I.; Phillips, C. B.; and Webber, B. L. 1993. *Simulating Humans*. New York: Oxford University Press.

Bateman, J. A.; and Paris, C. L. 1989. Phrasing a Text in Terms the User Can Understand. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*. San Francisco: Morgan Kaufmann Publishers.

Cassell, J.; Bickmore, T.; Campbell, L.; Vilhjálmsson, H.; and Yan, H. 2000. Human Conversation as a System Framework: Designing Embodied Conversational Agents. In *Embodied Conversational Agents,* ed. J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, 29–63. Cambridge: The MIT Press

Cassell, J.; Sullivan, J.; Prevost, S.; and Churchill, E., eds. 2000. *Embodied Conversational Agents*. Cambridge, MA: The MIT Press.

Cassell, J.; Vilhjálmsson, H.; and Bickmore, T. 2001. BEAT: The Behavior Expressive Animation Toolkit. In *Proceedings of the 28th Annual Conference on Computer Graphics,* Los Angeles, CA, 477–486. New York: Association for Computing Machinery.

Feng, D.; and Hovy, E. H. 2003. Semantics-Oriented Language Understanding with Automatic Adaptability. In *Proceedings of the 2003 International Conference on Natural Language Processing and Knowledge Engineering* (NLPKE), Beijing, China. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Fleischman, M.; and Hovy, E. H. 2002. Emotional Variation in Speech-Based Natural Language Generation. Paper presented at the International Natural Language Generation Conference, Arden House, NY, July 1–3.

Gratch, J. 2000. Émile: Marshalling Passions in Training and Education. Paper presented at the Fourth International Conference on Intelligent Agents, Barcelona, Spain.

Gratch, J.; and Marsella, S. 2004. A Domain Independent Framework for Modeling Emotion. *Journal of Cognitive Systems Research* 5(4): 269–306.

Hovy, E. H. 1990. Pragmatics and Natural Language Generation. *Artificial Intelligence* 43(2): 153–198.

Johnson, W. L.; Narayanan, S.; Whitney, R.; Das, R.; Bulut, M.; and LaBore, C. 2002. Limited Domain Synthesis of Expressive Military Speech for Animated Characters. Paper presented at the 7th International Conference on Spoken Language Processing, Denver, CO, September 16.

Johnson, W. L.; Rickel, J.; and Lester, J. C. 2000. Animated Pedagogical Agents: Face-to-Face Interaction in Interactive Learning Environments. *International Journal of AI in Education* 11: 47–78.

Larsson, S.; and Traum, D. 2000. Information State and Dialogue Management in the TRINDI Dialogue Move Engine Toolkit. *Natural Language Engineering* 6(3–4): 323–340.

Lazarus, R. 1991. *Emotion and Adaptation*. New York: Oxford University Press.

Lester, J. C.; Towns, S. G.; Callaway, C. B.; Voerman, J. L.; and FitzGerald, P. J. 2000. Deictic and Emotive Communication in Animated Pedagogical Agents. In *Embodied Conversational Agents,* ed. J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, 123–154. Cambridge, MA: The MIT Press.

Kendon, A. 1967. Some Functions of Gaze Direction in Social Interaction. *Acta Psychologica* 26(1): 22–63.

Marsella, S.; and Gratch, J. 2002. A Step toward Irrationality: Using Emotion to Change Belief. In *Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems*. New York: Association for Computing Machinery.

Marsella, S.; Gratch, J.; and Rickel, J. 2003. Expressive Behaviors for Virtual Worlds. In *Life-like Characters Tools, Affective Functions and Applications,* ed. H. Prendinger and M. Ishizuka. Berlin: Springer-Verlag.

Marsella, S.; Johnson, W. L.; and LaBore, C. 2000. Interactive Pedagogical Drama. In *Proceedings of the Fourth International Conference on Autonomous Agents.* New York: Association for Computing Machinery.

Matheson, C.; Poesio, M.; and Traum, D. 2000. Modeling Grounding and Discourse Obligations Using Update Rules. In *Proceedings of the First Conference of the North American Chapter of the Association for Computational Linguistics*. East Stroudsburg, PA: Association for Computational Linguistics.

Newell, A. 1990. *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.

Pelachaud, C.; Badler, N. I.; and Steedman, M. 1996. Generating Facial Expressions for Speech. *Cognitive Science* 20(1): 1–46.

Pellom, B. 2001. SONIC: The University of Colorado Continuous Speech Recognizer. University of Colorado, #TR-CSLR-2001-01, March, Boulder, CO.

Poggi, I.; and Pelachaud, C. 2000. Emotional Meaning and Expression in Performative Faces. In *Affective Interactions: Towards a New Generation of Computer Interfaces,* ed. A. Paiva, 182–195. Berlin: Springer-Verlag.

Rickel, J.; and Johnson, W. L. 2000. Task-Oriented Collaboration with Embodied Agents in Virtual Worlds. In *Embodied Conversational Agents,* ed. J. Cassell, J. Sullivan, S. Prevost, and E. Churchill. Cambridge, MA: The MIT Press.

Rickel, J.; Marsella, S.; Gratch, J.; Hill, R.; Traum, D.; and Swartout, W. 2002. Toward a New Generation of Virtual Humans for Interactive Experiences. *IEEE Intelligent Systems* 17(4)(July/August): 32–38.

Scherer, K. 1984. On the Nature and Function of Emotion: A Component Process Approach. In *Approaches to Emotion,* ed. K. R. Scherer and P. Ekman, 293–317. Hillsdale, NJ: Lawrence Erlbaum Associates.

Swartout, W.; Gratch, J.; Hill, R.; Hovy, E.; Lindheim, R.; Marsella, S.; Rickel, J.; and Traum, D. 2005. Simulation Meets Hollywood: Integrating Graphics, Sound, Story and Character for Immersive Simulation. In *Multimodal Intelligent Information Presentation*, ed. Oliviero Stock and Massimo Zancanaro. Berlin: Springer.

Thalmann, D. 1993). Human Modeling and Animation. In *Eurographics '93 State-of-the-Art Reports*. Geneva, Switzerland: European Association for Computer Graphics.

Traum, D. 2003. Semantics and Pragmatics of Questions and Answers for Dialogue Agents. In *Proceedings of the Fifth International Workshop on Computational Semantics*, 380–394. Dortrecht, The Netherlands: Kluwer.

Traum, D.; and Allen, J. F. 1994. Discourse Obligations in Dialogue Processing. In *Proceedings of the 32nd Annual Meeting of Association for Computational Linguistics*. East Stroudsburg, PA: Association for Computational Linguistics.

Traum, D.; and Rickel, J. 2002. Embodied Agents for Multi-Party Dialogue in Immersive Virtual Worlds. Paper presented at the First International Conference on Autonomous Agents and Multi-agent Systems, Bologna, Italy, July 15–19.

Traum, D.; Rickel, J.; Gratch, J.; and Marsella, S. 2003.

Negotiation over Tasks in Hybrid Human-Agent Teams for Simulation-Based Training. In *Proceedings of the Second International Joint Conference on Autonomous Agents and Multi-Agent Systems,* 441–448. New York: Association for Computing Machinery.

Traum, D.; Robinson, S.; and Stephan, J. 2004. Evaluation of Multi-Party Virtual Reality Dialogue Interaction. In *Proceedings, Language Resources and Evaluation Conference* (LREC 2004). Paris: Evaluations and Language Resources Distribution Agency.

Traum, D.; Swartout, W.; Marsella, S.; and Gratch, J. 2005. Fight, Flight or Negotiate: Believable Strategies for Conversing under Crisis. In *Proceedings of the Fifth International Working Conference on Intelligent Virtual Agents*. Lecture Notes in Computer Science. Volume 3661. Berlin: Springer-Verlag.

Wang, D.; and Narayanan, S. 2002. A Confidence-Score Based Unsupervised MAP Adaptation for Speech Recognition. Paper presented at the Proceedings of the 36th Asilomar Conference on Signals, Systems and Computers. November, Pacific Grove, CA.

William Swartout is the director of technology for USC's Institute for Creative Technologies and a research associate professor of computer science at USC. Swartout has been involved in the research and development of artificial intelligence systems for more than 30 years. His particular research interests include intelligent agents, immersive virtual reality, knowledge-based systems, knowledge representation, knowledge acquisition, and natural language generation. He received his Ph.D. and M.S. in computer science from MIT and his bachelor's degree from Stanford University. Swartout is a Fellow of the AAAI, has served on the AAAI Executive Council, and is past chair of the Special Interest Group on Artificial Intelligence (SIGART) of the Association for Computing Machinery (ACM).

Jonathan Gratch is a project leader at the University of Southern California's Institute for Creative Technologies and a research assistant professor in the computer science department. His research focuses on virtual humans (artificially intelligent agents embodied in a humanlike graphical body) and cognitive modeling. He studies the relationship between cognition and emotion, the cognitive processes underlying emotional responses, and the influence of emotion on decision making and physical behavior, and he has authored more than 100 technical articles and organized multiple international meetings on topics including virtual humans and emotion. He completed his Ph.D. in computer science at the University of Illinois in Urbana-Champaign and can be reached at gratch@ ict.usc.edu or www.ict.usc.edu/~gratch.

Randall W. Hill, Jr. is the director of applied research at the University of Southern California's Institute for Creative Technologies. He earned his B.S. degree from the United States Military Academy at West Point and his M.S. and Ph.D. degrees in computer science from the University of Southern California in 1987 and 1993, respectively.

Eduard Hovy leads the Natural Language Research Group at the Information Sciences Institute of the University of Southern California. He is also the deputy director of the Intelligent Systems Division, as well as a research associate professor of the computer science department of USC and an advisory professor of the Beijing University of Posts and Telecommunications. He completed a Ph.D. in computer science (artificial intelligence) at Yale University in 1987. His research focuses on information extraction, automated text summarization, question answering, the semiautomated construction of large lexicons and ontologies, machine translation, and digital government. Hovy regularly serves in an advisory capacity to funders of NLP research in the United States and European Union. He is the author or coeditor of five books and over 170 technical articles. URL: www.isi.edu/~hovy.html

Stacy Marsella is a project leader at the University of Southern California's Information Sciences Institute and a research assistant professor in the computer science department. His research interests include modeling human cognitive, emotional and social processes. He also explores the application of these models to the design of virtual humans for social training environments. Recent interests include research into the role that our beliefs about others play in our social interactions and research into modeling the dynamics of emotional responses. He completed his Ph.D. in computer science at Rutgers University. He can be reached at marsella@isi.edu and www.isi.edu/~marsella.

Jeff Rickel (1963–2003) was a project leader in the Intelligent Systems division of ISI and a research assistant professor at the USC School of Engineering department of computer science. He specialized in the design of robotic intelligent agents designed to serve as instructors for humans.

David Traum is a research scientist at the Institute for Creative Technology (ICT) and a research assistant professor of computer science at the University of Southern California. He completed his Ph.D. in computer science at University of Rochester in 1994. His research focuses on collaboration and dialogue communication between agents, including both human and artificial agents.

AAAI Annual Business Meeting

The annual business meeting of the American Association for Artificial Intelligence will be held at 12:45 PM, Monday, July 17, 2006 in the Seaport World Trade Center Complex in Boston, Massachusetts.