Design criteria, techniques and case studies for creating and evaluating interactive experiences for virtual humans

Jonathan Gratch, Stacy Marsella, Arjan Egges, Anton Eliëns, Katherine Isbister, Ana Paiva, Thomas Rist, Paul ten Hagen[1]

How does one go about designing a human? With the rise in recent years of virtual humans this is no longer purely a philosophical question. Virtual humans are intelligent agents with a body, often a human-like graphical body, that interact verbally and non-verbally with human users on a variety of tasks and applications. At a recent meeting on this subject, the above authors participated in a several day discussion on the question of virtual human design. Our working group approached this question from the perspective of interactivity. Specifically, how can one design effective interactive experiences involving a virtual human, and what constraints does this goal place on the form and function of an embodied conversational agent. Our group grappled with several related questions: What ideals should designers aspire to, what sources of theory and data will best lead to this goal and what methodologies can inform and validate the design process? This article summarizes our output and suggests a specific framework, borrowed from interactive media design, as a vehicle for advancing the state of interactive experiences with virtual humans.

"What a piece of work is man!!"

At the end of the day, what *should* a virtual human be like? It is seductive to hold up ourselves as the ideal for virtual humans; however there are some obvious complications with this view. Seemingly trivial concerns about virtual humans getting bored or going on strike point to a deeper tension underlying our field. On the one hand, there are obvious reasons for understanding, modeling, and in some sense mimicking human behavior. On the other hand, virtual humans are a tool that must efficiently fulfill a role in an overall system and their design characteristics must be subordinate to the overall goals of this system. If the goal in creating a virtual human is to support "effective" interaction in some sense of that term, it is natural to clarify this tension and understand what implications it has for the choice of theoretical models and sources of data to inform virtual human design.

There is a strong argument for studying human-to-human interaction as it occurs "in nature" as a basis for virtual human design, which we will call a *human-ecological perspective*.[2] The intuition is that people have evolved and developed within an "ecological niche" that emphasizes face-to-face communication with other people; that we are very skilled in such settings; and if only computers could acquire this human skill, human-computer interaction would be more effective and efficient. Followers of this perspective look to psychology and linguistics as a source of

[1] Jonathan Gratch and Stacy Marsella are at the University of Southern California, Arjan Egges at MIRALab, Anton Eliëns, at the Vrije Universiteit Amsterdam, Katherine Isbister at Rensselaer Polytechnic Institute, Ana Paiva at INESC-ID and IST, Tagus Park, Thomas Rist at the University of Augsburg and Paul ten Hagen at Epictoid BV.

[2] The term ecological is intended to emphasize the collection of data in naturalistic settings, but it also carries the subtext that these are stable, ubiquitous settings that have some constancy across the individual's development and potentially their evolutionary history. James J. Gibson coined the term "ecological psychology" to emphasize that, through coevolution of organisms with their environment and through lifelong experience with stable ecological niches, organisms develop certain patterns of interaction that they will carry forward into novel situations. This is also related to the notion of ecological validity in experimental design (Schmuckler, 2001), whereby experiments may uncover findings that never arise in practice if the stimuli are too divorced from the people's everyday experience.

Dagstuhl Seminar Proceedings 04121
Evaluating Embodied Conversational Agents
http://drops.dagstuhl.de/opus/volltexte/2006/462

theory and data to narrow this gap, which is wide indeed. Unlike text based interactions, face-to-face interactions involve a rich exchange of information across several modalities including rapid feedback that allows the participants to (in comparison to text) rapidly converge on a shared understanding. By studying human-to-human interaction, psychologists and linguists have posited a number of functions that underlie conversation (turn taking, repair, grounding) and the mapping between these functions and surface behavior. People bring these skills and perceived expectations when confronting a human-appearing graphical entity and, the argument proceeds, by meeting and leveraging these expectations, virtual humans can promote more efficient communication than more traditional human-computer interfaces.

There are some complications with the ecological perspective. Of course, there are the obvious technical and theoretical challenges that accurately understanding and modeling human in its full richness (and in truth, addressing these challenges is an end in itself for many virtual human researchers). A more subtle issue is this perspective can promote a conservative view of human interaction. By the nature of psychological study, human-to-human research tends to exhaustively explore a few canonical naturalistic settings. From the perspective of a virtual human researcher (who tends to view this literature from the outside and is motivated to find "the top 10 rules" that will make their system work), this can create a tendency to unduly elevate particular interaction settings and obscure the enormous variability and adaptability human-to-human interaction. The norms and styles of interaction vary enormously depending on the setting, the power relationship between participants, the artifacts in the environment, etc., and people readily evolve their norms, symbols and expectations when confronted with new situations and new medium. Many envisioned virtual human applications involve interactions not considered by the psychology community, and where people would not necessarily follow the conventions of any particular ecological study. Finally, given the pressure to develop (and receive funding for) real applications, the ecological perspective may be unnecessarily complex as a starting point (just because people do it, doesn't mean they need to) and unnecessarily limiting in terms of the range of techniques and data that could be applied to the problem.

An alternative view is to take a *media perspective,* emphasizing people's adaptability to new interaction styles and drawing on interactive media, computer games, the arts, and filmmaking as a source of theory and data. The essence of this view is that, although ecological interactions between people provide a good starting point, we can do much better in terms of promoting efficient interaction.[3] "Media" has developed stylized presentation styles that people also have considerable experience interacting with that (arguably) have greater communicative efficiency that face-to-face interaction. For example, in a movie, one must rapidly convey a person's life history, personal dispositions and emotions. Actors and animators adopted a style that departs markedly from naturalistic human behavior: there is a strong tendency toward stereotypical appearances and mental state is much more transparently conveyed through behavior and expression (Coats, Feldman, & Philippot, 1999). These "tricks" extend well beyond the character's behavior, and media practitioners use a range of contextual cues to promote the observer's rapid understanding, including the choice of genre, backstory, staging, cinematography, and sound tracks. Consumers of such media have learned to accept these "unnatural" conventions and readily exploit them to improve their understanding of an interaction.

This view comes with its own set of complications and limitations. Chief among these is the fact that these media are not interactive in the sense envisioned by virtual human researchers so there

[3] This perspective is not necessarily un-ecological, given that many American children watch television for more than forty hours a week. Thus, a 'media ecology' may increasing

is no real data about their applicability in this context. Further, with a few exceptions, the findings are anecdotal and the "theory" underlying an effective performance consists of tacit knowledge in the head of the actor or animator. It is very difficult to extract useful rules of thumb to guide virtual human design (although the same may be said for the human-ecological perspective, but in this case due to the complexity of the findings rather than their opacity). It is also unclear the extent to which people can easily learn these artificial conventions, and there is evidence that such conventions are more stable and difficult to acquire (and in this sense more ecological) than one might think. For example, many of our older citizens seems less comfortable with some of the more recent forms of mediated interaction (email, instant messaging, etc.) and there is some evidence that certain forms of interaction, for example, extensive interaction with so-called "twitch" games can cause persistent changes in how people interact with their environment (Green & Baveller, 2003).

A broader notion of ecology can reconcile these two perspectives into what might be called an *interaction-ecological* perspective. From this perspective, people can fluidly transition between different interaction contexts (ecological niches), some human-to-human, others mediated, to which they bring different, though stable, skills, norms and expectations.[4] This view retains the insight that a virtual human designer can promote efficient interacting by cuing and leveraging off of existing norms and skills, but it opens up a broader pallet with which to accomplish this goal. This view can draw on naturalistic research to inform our understanding of the underlying form and function of interaction, but broadens the notion of ecology to encompass the mediated interactions that occur in everyday life. This view:

- recognizes that people readily respond to "unnatural" stimuli, including stylized "supernormal stimuli"**—**in the sense that a red pencil is a more effective elicitor of feeding behavior in baby Herring Gulls than their mother's actual beak (Tinbergen & Perdeck, 1950)**—**as well as learned artistic conventions**—**such as Tex Avery's popping eyeballs**—**and that the arts as well as psychology have a role in articulate what these stimuli may be;
- accepts that what is natural depends on the context (ecological niche); that a variety of factors (appearance, backstory, prior training, etc.) can impact what context the user believes himself to be in; and that context must be kept in mind, both when eliciting human-to-human interaction data and when considering the staging, appearance, etc. of the virtual human application;
- emphasizes that people can adapt, and the norms of interaction with a specific context need not be adhered to slavishly, although some aspects on an interaction are probably more malleable than others**—**for example, in our own work, people seem to eventually adapt to the slower pace of virtual human conversations, but they have more difficulty recognizing (as they are gazed upon by the virtual human's sightless eyes) that their own gestures are meaningless to the character; and

Ultimately, interaction is a negotiation. Part of this negotiation is played out between a virtual human and a user, but part of it will be played out between the virtual human research community and our user base. Where the two sides will meet is an open question.

Affordances for interaction

With this prelude, our working group focused on a particular design methodology that could inform this functional perceptive on interactivity. In this we adapted a standard methodology used for the design and evaluation of interactive media based on the notion of *affordances (Gaver, 1996; Gibson, 1979; Norman, 1990).* Affordances derive from an ecological perspective

[4] This also relates to the notion of *lifeworlds* (Agre & Horswill), which emphasizes that several ecological niches may share the same physical space depending on the organisms' goals, percepts and actions.

on interaction. They refer to recognized properties of objects that signal a particular way of interacting (e.g., a handle affords grasping and a table affords support). In Gibson's original formulation, affordances were assumed to reflect evolved brain structure. More recent theorists have generalized the term to encompass experientially acquired and context-specific mappings, and this more recent perspective, consistent with the interaction-ecological perspective sketched above, is these sense we adopt.

In the present context, the overarching intuition is that 1) certain features of a virtual human and its setting (both visual and behavioral) will cue a user to interact in a certain fashion; 2) certain pre-experience manipulations (prior training, backstory, etc.) can impact the salience of these cues; and 3) the choice of these prior manipulations and cues, collectively, will be more or less successful in promoting effective interaction. We then applied this design method retrospectively to several existing virtual human applications, identifying a number of features that both contributed to, and detracted from effective interaction.

We took as a starting point several basic assumptions. First, interaction is central: a virtual human's primary function is to promote effective human-computer interaction, both directly with the virtual human but indirectly with some larger application (though our emphasis here is on the virtual human). Second, that interaction can be usefully analyzed as a graph structure that represents *what* a user can do. Thus, the arcs in the graph are the interaction "moves" that a user can make (dialogue, mouse clicks, etc.) and the nodes are (suitably abstracted) states of the virtual human. Note that we are not assuming that the system directly represents this graph, but rather that this graph is a useful analytic abstraction of the interaction. Third, that "effective" interaction can be defined as some *policy* over this graph (e.g., the user must reach some end state subject to constraints, or the user must follow a certain trajectory in the graph). Fourth, that various features and contextual factors embedded in the system can be usefully characterized as *affordances* to act, that cue the user (more or less successfully) towards obeying this policy.

If one can cast a system as such a graph, policy and collection of cues, there are a variety of assessment criteria one can apply to the system. One obvious set of tests relates to the properties of the graph. Is the end state reachable? Are desirable paths connected? There are also a variety of ways to assess affordances. For example, do they cue the user to illegal moves (moves that lead outside the graph) or undesirable moves (moves that conflict with the policy)? We also spoke of generalizing the notion away from a narrow interpretation as a tendency to act. An affordance is a relational concept that links features of the world with the intentions, perceptions, and capabilities of the user. It is the user's *interpretation* of cues and rather than simply focusing on the ultimate act, an affordance can influence the antecedents of action. Thus we could speak of, and potentially measure through self report, affordances of intrinsic motivation, affordances of control, affordances of novelty, etc. These perceived senses will act as mediating variables that indirectly influence a users choice of action. For example, they will only continue to interact if they are motivated, they will only consider actions that they perceive to have some control/influence over the interaction, etc.

Case Studies

The primary contribution of our working group was the detailed walkthrough of virtual human applications, identifying, to the best of our ability, the interaction graph, the intended policy, the affordances provided by the virtual human, and the extent to which the affordances were recognized and cued the appropriate action. We considered several systems. These included a traditional computer game ("the curse of monkey island") where the user interacts with a mouse and multiple-choice dialogue moves, a system where the virtual human plays the role of a teammate in a training exercise (the University of Southern California's Mission Rehearsal

Exercise) where the user interacts through spoken dialogue (Rickel et al., 2002), a system where user's control the emotional state of a virtual human through a tactile interface (Prada, Vala, Paiva, Höök, & Bullock, 2003), and a system where the virtual human plays the role of a presentation agent (RUDY).

We spent the most time discussing the Mission Rehearsal Exercise. This system allows a user to play the role of a U.S. Army lieutenant on a peacekeeping mission in Bosnia. They interact with multiple virtual humans in the scenario through spoken voice and virtual humans can respond with voice and gesture. Characters are projected on a 2.5x10 meter screen and appear life-sized to the user. We analyzed a video segment of U.S. Army cadet interacting with the system (www.ict.usc.edu/~gratch/media/8-03-westpoint3.mov). Several findings came out of this analysis (some that were not immediately obvious to the developers of the system). Here we illustrate the basic analysis:

What is the (implicit) graph and/or space?

- Dialog moves: The user can ask information about the state of the world, can elicit suggested actions, can give orders. An explicit task model (which the user is expected to have some familiarity with) constrains the interaction to a small number (40) possible states and small number (60) of possible actions.
    -

What is the desirable path through this space?

- There is a "good" outcome and a "bad" outcome based on the user's choice of actions
- There is an implicit goal to have the user emotionally aroused. Simulation events should increase their uncertainty and decrease their perceived sense of control over the ultimate outcome
- There is an implicit goal to use the immersive environment to improve recall
- There are local rewards that improve intrinsic motivation. Getting the sergeant to understand you is similar to the frustration of getting off a level of a game. There is some question as to if this communication problem "part of the fun?" (as exploited in certain games) or does this run contrary to the goal of creating an immersive experience

What techniques are used to cue user toward desirable path

- The agent has local repair moves to guide the user back to the proper path (e.g. "I can't understand you")
- The role assigned to the user cues them as to their possible actions (give orders to subordinates, follow orders of superiors).

Where did the guidance fail?

- Certain dialogue failure feedback cued the wrong action. An utterance "I can't hear you, sir." Was added so that the user would repeat their utterance, however it cued them to speak louder and slower, hindering recognition ability.

Future Work

Our working group anticipates holding future discussions on this topic. Several questions remain. It is unclear if such a simplistic notion of an interaction graph will be workable. From an analysis perspective, we may wish to consider both a graph that characterizes the virtual human as it is (truth) vs. the graph as the user perceives it (perception), and look to discrepancies in informing the system design. Complementary to analyzing the system as a graph, perhaps the system should maintain an explicit characterization as the user as a graph with some policy, and use this in informing its interaction strategy. It may be more appropriate to construct a frame of analysis that encompasses both user and ECA in a single interaction graph. It is also unclear if it is appropriate to think in terms of a fixed graph that the user (perhaps) comes to recognize, or if the graph itself a dynamic structure. We also need to further research how this formalism has been applied in other design contexts and to go into greater depth.

References

Agre, P., & Horswill, I. Lifeworld Analysis.

Coats, E. J., Feldman, R. S., & Philippot, P. (1999). The influence of television on children's nonverbal behavior. In P. Philippot, R. S. Feldman & E. J. Coats (Eds.), *The social context of nonverbal behavior* (pp. 156-181). Paris: Cambridge University Press.

Gaver, W. W. (1996). Affordances for interaction: the social is material for design. *Ecological Psychology, 8*(2), 111-129.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston: Houghton Mifflin.

Green, S. C., & Baveller, D. (2003). Action video game modifies visual selective attention. *Nature, 423*.

Norman, D. A. (1990). *The design of everyday things*. Cambridge, MA: MIT Press.

Prada, R., Vala, M., Paiva, A., Höök, K., & Bullock, A. (2003). *FantasyA - The Duel of Emotions.* Paper presented at the the 4th International Working Conference on Intelligent Virtual Agents, Kloster Irsee, Germany.

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

Schmuckler, M. A. (2001). What is ecological validity? A dimensional analysis. *Infancy, 2*(4), 419-436.

Tinbergen, N., & Perdeck. (1950). On the Stimulus Situation Releasing the Begging Response in the Newly Hatched Herring Gull Chick (Larus argentatus argentatus Pont). *Behavior, 3*, 1-39.