Design and Evaluation of Embodied Conversational Agents: A Proposed Taxonomy

Katherine Isbister, Ph.D.
1904 23rd Street
San Francisco, CA 94107
(+1)415-722-1945
ki@katherineinterface.com

Patrick Doyle
Stanford University
Gates Computer Science Bldg. 2A
Stanford, CA 94305-9020
(+1) 650-723-6707
pdoyle@cs.stanford.edu

ABSTRACT
This workshop call demonstrates that our field is eager to move beyond first-generation generalist projects, toward a more mature practice. To do so, we seek to set up a common set of expectations and criteria for how to judge our work. In this paper, we propose some subclasses of embodied conversational character research and design, with criteria for describing and evaluating research and design advances in each. We suggest that researchers in this field could benefit from carefully identifying their own areas of expertise and contribution, and then looking for ways to collaborate on standards and share advances within these sub-areas. Presenting results, then, would require making clear the sub-areas addressed by the particular project, with evaluations appropriate to those areas included. We believe this approach can help the research community to clarify contributions, and more easily build a common base of knowledge.

Categories and Subject Descriptors
I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence; D.2.8 [**Software Engineering**]: Metrics

General Terms
Design, Languages, Human Factors.

Keywords
Conversational agents, design categories, evaluation criteria.

1. INTRODUCTION
The effort to create an embodied conversational agent is, by its very nature, multidisciplinary. Creating a fully realized agent requires the application of diverse disciplines ranging from agent systems [4,5,6,13,17,22,28,29], models of emotion [7,8,27], graphics [2,25] and interface design [18,30], to sociology and psychology [9,16,23,24,35], and even art, drama, and animation techniques [17, 33]. The practitioners of these disciplines do not share a common language, even when describing components of the common goal; the criteria for critical evaluation, when they exist at all, vary wildly among these disciplines, and there is no common objective measure by which we can determine whether a research product is "good."

Often, our research papers describe the construction of a complete prototype, mixing discussions of technical innovations with new application areas and interaction techniques [1,2,14,15,32,34]. Choices about the appearance, personality, and behaviors of the agent are frequently made on the basis of an introspective examination of personal preferences, and in many cases do not accurately reflect the goals of the design or the qualities of the audience with whom the agent is ultimately intended to interact. Rigorous evaluations of benefits to the user (e.g., [20]) are rare, and even when performed are subject to considerable criticism owing to the difficulty of finding objective measures of success.

Our intent in this paper is not to criticize past work. On the contrary, these failings are not the result of flawed research but the necessary compromises made in the exploration of a new research area, and one in which nearly every major architectural or design decision is dependent upon a combination of factors springing from widely different bodies of knowledge. However, to continue to make best progress, we will have to develop a set of criteria for design and for evaluation that makes use of all of these disciplines. Though we need to work together to create successful characters, we need to preserve the standards for excellence and methods for extensibility that each specialty brings to our work.

The purpose of this paper is to attempt a broad taxonomy of the research areas contributing to the creation of embodied conversational agents. Although these agents are the collective goal, each area is making a different functional contribution and has distinct methods and measures for evaluating work. Our hope is twofold: first, to encourage a recognition of these divisions, making it possible for researchers to be more clear about what their work is, and is not, attempting to accomplish, and second, to provide a better basis for understanding how to evaluate agents that only implement solutions to some parts of the entire problem.

2. THE TAXONOMY
In order to motivate our taxonomy, we follow the approach of Franklin and Graesser [11] by examining several standard definitions for conversational agents, intelligent characters, believable agents, etc.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
AAMAS '02, July 15-19, 2002, Bologna, Italy.
Copyright 2002 ACM 1-58113-000-0/00/0000…$5.00.

Table 1. Major categories of embodied conversational agent research

Category	Examples of Properties	Related Disciplines	Criteria for Success	Evaluation Techniques
Believability	Emotion, strong personalities, variability in movement and response, personalizability, idiosyncracy, interesting, appearance of goals, appearance of caring what happens	Art, animation, film, literature, cognitive psychology, neurobiology	Agent conveys the "illusion of life" to the viewer/user.	Subjective; does the viewer find the agent's behavior believable? Use of audience survey-style measures.
Social interface	Social context, social behaviors, knowledge of other agents, empathy, dialog, ability to cooperate	Cognitive and social psychology, sociology	User is able to interact in an intuitive and natural way with the agent to perform intended task.	Qualitative measures from user of agent's friendliness, helpfulness, or intuitive communication ability; quantitative measures of speed, ease, satisfaction with achievement of task.
Application domains	Domain knowledge, contextuality, timeliness, effectiveness, risk/trust	Disciplines appropriate to the application domain and to the task; interaction design, psychology, sociology	Agent achieves the goals of the application (e.g., training).	Varied, depending on the application and goals. Did the agent achieve the goal? Objective measures of performance.
Agency and computational issues	Autonomy, responsiveness, reactivity, reliability, completeness, efficiency, goal-directedness, optimality	Computer science, philosophy	Elegance of system, parsimony, speed, selection of optimal actions, proofs guaranteeing certain behaviors.	Successful operation of the agent in "real-world" domains according to criteria of speed, efficacy, reliability, error handling, etc.
Production	Professional and consistent quality in final visuals, dialog, behavior, and interaction mechanisms, and the integration of the whole.	Professional design, production, and project management; systems integration.	User's experience is not marred by lack of quality or inconsistency. Team's experience in creating character was efficient and positive.	User evaluation/ranking of production values and smoothness of overall experience. Confirmation that the character is 'read' as was intended. Evaluation of effectiveness of production techniques on efficiency and team satisfaction with process.

Loyall writes [21]: "Believable agents are personality-rich autonomous agents with the powerful properties of characters from the arts." A believable agent has personality, emotion, self-motivation, change, social relationships, consistency, and presents the illusion of life (appearance of goals, concurrent pursuit of goals, parallel action, reactive and responsive, situated, resource bounded, exist in a social context, broadly capable, well-integrated capabilities and behaviors).

According to Blumberg, "…an autonomous animated creature is an animated object capable of goal-directed and time-varying behavior." [5] A creature must react, be seen as having an independent existence, have choices to make, reveal its intentionality, appear to care what happens to it, adapt, and display variability in movement and response.

Reilly says that believable agents are, "autonomous, interactive agents that have the qualities that have made the non-interactive characters of traditional media believable." [28] They may not be intelligent or realistic, but they will have strong personalities.

Hayes-Roth and Doyle claim that "animate characters" redefine traditional agent design problems [12]. In addition to possessing empathy, personality, and a capacity for social relations, their behaviors must be variable rather than reliable, idiosyncratic instead of predictable, appropriate rather than correct, effective instead of complete, interesting rather than efficient, and distinctively individual as opposed to optimal.

Stone and Lester [32] describe animated pedagogical agents as possessing three key properties: timely domain coverage (that is, coverage of the educational topics), contextuality (appropriate explanations for the situation), and continuity (coherent behaviors, pedagogically and believably).

Perlin and Goldberg are concerned with building believable characters "that respond to users and to each other in real-time, with consistent personalities, properly changing moods and without mechanical repetition, while always maintaining an author's goals and intentions." [25]

Trappl and Petta describe synthetic characters as needing different abilities in different contexts; "animators might…be assisted in

the…delicate and ephemeral task of ensuring consistent and believable patterns of behaviour… in yet other settings, achieving e.g. various degrees of agent autonomy can play an essential role in providing effective assistance to users…" [34]

In characterizing believable agents, Bates requires "only that they not be clearly stupid or unreal." Such broad, shallow agents must "exhibit some signs of internal goals, reactivity, emotion, natural language ability, and knowledge of agents…as well as of the…micro-world." [3]

For Foner [10], the critical issues are autonomy, personalizability, discourse (ability to carry on two-way dialog), risk/trust, domain (appropriate domain for the agent), graceful degradation, ability to cooperate, anthropomorphism, and choosing a setting in which users' expectations for the agent can be met.

Considering these definitions, several themes emerge. First are the properties of traditional agents – autonomy, responsiveness, reactivity, situatedness, and goals. Next comes the ability to function socially; the agent should be able to carry on discourse, understand other agents, be able to cooperate, have a capacity for social relationships and social behaviors. Third is the agent's capacity for behaving believably. It should exhibit graceful degradation, interesting variability, idiosyncrasy, personalizability, contextuality, appropriate behaviors, intentionality, and should do nothing "clearly stupid or unreal." Fourth are its particular domain behaviors – the agent should behave contextually, cover domain topics in a timely way (for a pedagogical agent), it should be effective, and have well-integrated capabilities with the environment.

Each of these areas has a large body of knowledge, often stemming from the attempt to understand comparable qualities of human beings. However, integrating these disciplines (e.g. agency with believability) has required combining areas with very different knowledge and goals, and whose criteria for evaluation range from theoretically provable, to statistically measurable, to purely subjective. Our claim is that we must be more precise as researchers about which elements our work focuses on and which we are only approximating, so that we use appropriate measures when examining new contributions.

Table 1 summarizes the four major areas we have drawn out of these definitions, together with a fifth, Production, that encapsulates the problems of integrating work done in the other four. In the following sections we will provide more details to clarify what each category is, what its contributions are, and how work in that category should be evaluated.

2.1 Appearance and behavioral believability

This specialty is concerned with making a person's visceral reaction to the character a powerful one—to evoke the 'illusion of life'. This would include appearance as well as sound and movement—all things that make a character more sensorially engaging. Research in this area is not limited to "realism," but also includes media goals for exaggerating and enhancing human reactions to the character.

Some examples: researchers who enhance the realism of character walks and movements; those who can specify and create the right visual appearance and style for specific character applications, working within technical tradeoffs; researchers who create appealing and natural-sounding speech.

Skills required: training in the media being used, and in the observation of the relevant qualities of human beings and their perceptual systems.

Criteria for success: Some form of response from end users that the character quality being produced is "lifelike" or "larger than life" in the appropriate way. Bates writes, "To our knowledge, whether an agent's behavior produces a successful suspension of disbelief can be determined only empirically." [3]

Evaluation techniques: Commercial specialists rely on audience satisfaction measures, within the context of a polished final experience by the right target audience in the right setting. Practitioners in schools also use critique by other qualified specialists, to help evaluate success, when audience surveys are not practical.

2.2 Social interface techniques

This specialty innovates and enhances the manner in which people interact with embodied conversational agents. This includes examining the pros and cons of various input techniques depending upon situation, exploring the range of social roles and interaction styles agents can inhabit, and distilling principles around these that can be used across agent projects. The focus is on how the user engages with the agent, and what works and doesn't work about any given technique (as opposed to a focus on the construction of new input or interaction methods). This category also includes the design of social interaction between multiple agents and the user.

Some examples: Designing characters that use gestures in communication, and evaluating the effectiveness of these gestures in aiding communication. Studying the effect of modality consistency (I type, character types; versus I type, character talks) on social engagement and trust. Examining the effectiveness of role-appropriate small talk in eliciting role-appropriate responses from users.

Skills required: training in the human interaction strategies being examined (e.g. nonverbal communication, social roles, etc.). Training in the evaluation and iteration of these strategies, within the mediated context.

Criteria for success: Some form of qualitative and quantifiable response from end users that the character interaction being produced is engaging, helpful, and/or intuitive in the manner that would be predicted by application of the relevant social tactics.

Evaluation techniques: Surveys and rigorous observation of end users in the right target audiences for the interaction at hand. When possible, use of experimental methodology to isolate results to the social interface strategy being used.

2.3 Application domains

This specialty seeks to establish what applications agents have that are useful to particular groups, and why. The focus is on thoroughly researching the application domain, and testing the completed embodied conversational character with real users in that target group, using meaningful benchmarks.

Some examples: A tutoring agent created by systematically observing human tutors, then testing the final character out with real students. A consumer assistance agent created after researching live customer support practices, tested in a real

customer service situation, using quantitative metrics for success (such as reduced live customer support volume).

Skills required: domain knowledge about the application area; design skill to apply this knowledge; user research skills to evaluate the success of the design within the domain context.

Criteria for success: production of a successful final character that achieves goals set for this application area (e.g. increase learning, decrease customer email). Role usefulness as perceived by domain experts as well as end users.

Evaluation techniques: Setting key benchmarks at the beginning, during research of the application area. Rigorous qualitative and quantitative analysis of performance against these benchmarks in the application domain, with the right user group.

2.4 Agency and computational issues

This specialty innovates computational techniques for creating successful embodied conversational characters. This specialty also researches issues involved with integrating various components into workable systems.

Some examples: Researchers who design architectures that can handle delivery of the synchronized multi-modal actions and reactions of a character; researchers who computationally model user input patterns such as eye gaze; researchers who generate frameworks for generating appropriate arcs of emotion in characters; researchers who create annotation frameworks for specifying character behavior.

Skills required: programming and system architecture skills, knowledge of the technological constraints of the systems being used.

Criteria for success: Parsimony, elegance and broad applicability of solutions; ability to deliver agreed-upon benchmarks of behavior/output (as set by both the computational specialist and the other specialists on the team).

Evaluation techniques: Evaluating code and the usefulness of a programming or architecture decision or standard relies on peer critique by others in the same knowledge area. Evaluation of the end result of the computational solution as manifested in character behavior should include the types of evaluation listed in appearance, interaction, and applications.

2.5 Production

This specialty is not often specifically addressed in our community's papers, yet it is an essential component in creating successful embodied conversational characters. By production, we mean techniques used to assure high quality overall production values in character looks and behaviors, and smooth integration and performance of all components. As the technologies that we use shift, so do best practices for achieving the level of professionalism necessary to elicit the user responses we seek.

Some examples: Contributions in this area could include sharing successful tactics for mapping and gathering the resources (people, hardware, software) needed to complete a project, creating schemes for file handling and asset processing, managing cycles of iteration and user testing, and quality testing before release. Providing information for other practitioners about potential pitfalls and best practices for using new technologies (e.g. input devices) would also be a valuable contribution in this area. Since many projects involve collaboration between multiple locations, best practices papers on achieving good results given this context would also be very useful.

Skills required: project management; ability to evaluate and distill best practices from project experiences.

Criteria for success: Demonstrated high level of quality in final character, as confirmed by fellow practitioners as well as target end users, achieved by the use of the production best practices described.

Evaluation techniques: Qualitative and quantitative end user measures of satisfaction with character, and its level of professional quality. Empirical comparisons of production tactics described with less effective practices, preferably with measurable benchmarks (requires less time, money, or people, measurably more positive experience for team members, etc.).

3. APPLYING THE TAXONOMY

We propose that researchers in our community might use this taxonomy in several ways.

1. *Use it to clarify and communicate primary skills.*

Identify one's area of expertise and deep knowledge, and make this known to the community.

2. *Use it to assemble appropriate teams, during project planning.*

Before beginning a project, list the competencies needed to achieve a successful enough user experience to address your primary research question. Then, connect with other researchers with the needed complementary skills, and design an overarching project that meets research goals for everyone. Alternatively, get permission to re-use a component from another researcher, acknowledging that component's creators.

3. *Use it to set evaluation benchmarks.*

Each interdisciplinary research group should set evaluation benchmarks and plans for each sub-area they plan to make a contribution in, relying on the evaluation expertise of each specialist.

4. *Use it to contextualize work for others in our community.*

When reporting results, make it clear where the primary goals and contributions lie, and remind the audience of the appropriate standards of evaluation.

The research community can help to bolster this approach by setting different standards of evaluation for each type of contribution. We should expect rigorous, contextual testing of anything that claims to address a real application need. We should expect empirical evaluation by appropriate target audiences of any advance in appearance technique. We should expect peer reviewable descriptions of any new architecture or computational technique, and if it claims to address a real interaction need as well, accompanying user evaluation of the success of the manifestation of that technique. We should allow for research papers that are case studies about production best practices.

We may also want to think about rewarding the reuse of components that others have contributed, when a researcher's primary goal for contribution does not involve that specialty. This might help to encourage the creation of extensible, reusable components, as they would show up again and again and get cited in the community as a whole. It also might release researchers

from feeling they need to wholly reinvent the wheel to gain attention for their project.

4. FURTHER CLASSIFICATION

The taxonomy we have described here is quite abstract and makes only broad distinctions between communities of practice (such as between those working on computational models of agency and those working on social interface questions). There is ample room for refinement.

The most obvious area for further subdivision is believability. Within that category there are many areas to draw upon for inspiration, as is suggested by Table 1: traditional film studies, the copious animation literature, motion studies, drama and acting, literature and writing, etc.

Another kind of subdivision can be made along the philosophical approach to agency and believability, which we have not touched on here. Most of the groups mentioned take the dramatic approach of determining what behaviors are recognizably lifelike and then producing agents that exhibit those behaviors in such a way as to maximize their believability. These agents are comparable to actors who consciously think about how to convey certain emotions or expressions to an audience, without necessarily experiencing the stimuli that provoke them. The alternative approach (exemplified in, e.g., [REF Blumberg's Thesis]) is to create agents that simulate natural systems and whose lifelike behavior arises from ethological considerations, e.g., the agent slowly becomes hungry over time and will therefore go in search of food, with its stomach growling. Given that the inspirations for these two lines of work are so dissimilar, it would be beneficial if we could quickly determine which is being drawn upon.

5. CONCLUSIONS

In this paper, we've outlined a taxonomy of specialties in the embodied conversational character research community. Each consists of a particular research agenda, set of skills, criteria for success, and evaluation techniques. We suggest that our community can set clearer and better benchmarks and create more extensible solutions by clarifying which specialties each project addresses, and by holding the project results to the standards of that specialty.

6. BIBLIOGRAPHY

[1] Andre, E. (ed.). Notes of the IJCAI-97 Symposium on Animated Interface Agents: Making Them Intelligent (Nagoya, Japan, August 1997).

[2] Badler, N. Real-time virtual humans. In Proceedings of 1997 Pacific Graphics Conference (Seoul, Korea, 1997).

[3] Bates, J. The nature of characters in interactive worlds and the Oz project. Technical report CMU-CS-92-200, School of Computer Science, Carnegie Mellon University, Pittsburgh PA, October 1992.

[4] Bates, J., Loyall, A. B., and Reilly, W. S. An architecture for action, emotion, and social behavior. Technical Report CMU-CS-92-142, School of Computer Science, Carnegie Mellon University, Pittsburgh PA, 1992.

[5] Blumberg, B. Old Tricks, New Dogs: Ethology and Interactive Creatures. Ph.D. Thesis, Media Lab., Massachusetts Institute of Technology, Cambridge MA, 1996.

[6] Cassell, J., Bickmore, T., Vilhjálmsson, H., Yan, H. More than Just a Pretty Face: Affordances of Embodiment. In Proceedings of CHI 2000, 52-59.

[7] Elliott, C. The Affective Reasoner: A Process Model of Emotions in a Multi-Agent System. Ph.D. Thesis, The Institute for the Learning Sciences, Northwestern University, 1992.

[8] Elliott, C., Lester, J., and Rickel, J. Integrating Affective Computing into Animated Tutoring Agents. In Notes of the IJCAI '97 Workshop on Animated Interface Agents: Making Them Intelligent (Nagoya, Japan, August 1997), 113-121.

[9] Fiske S. and Taylor, S. Social Cognition. McGraw-Hill, New York, 1991.

[10] Foner, L. What's an agent, anyway? A sociological case study. Agents Memo 93-01, Agents Group, Media Lab., Massachusetts Institute of Technology, Boston MA, May 1993.

[11] Franklin, S., and Graesser, A. Is it an agent, or just a program? A taxonomy for autonomous agents. In Agent Theories, Architectures, and Languages. Springer-Verlag, Berlin, 1996, 21-95.

[12] Hayes-Roth, B., and Doyle, P. Animate characters. Autonomous Agents and Multi-Agent Systems 1, 195-230.

[13] Hayes-Roth, B., Brownston, L., Huard, R., van Gent, R., and Sincoff, E. Directed improvisation. Technical Report KSL-94-61, Knowledge Systems Lab., Stanford University, Stanford CA, Sept. 1994.

[14] Hayes-Roth, B., van Gent, R., and Huber, D. Acting in character. In Creating Personalities for Synthetic Actors, Trappl, R., and Petta, P. (eds.) Springer-Verlag, Berlin, 1997.

[15] Isbister, K., Nakanishi, H., Ishida, T., and Nass, C. Helper Agent: Designing an assistant for human-human interaction in a virtual meeting space. In Proceedings of CHI 2000 (the Hague, Netherlands), 57-64.

[16] Isbister, K., and Nass, C. Consistency of personality in interactive characters: Verbal cues, non-verbal cues, and user characteristics. In International Journal of Human-Computer Studies 53(2), 251-267

[17] Lasseter, J. Principles of traditional animation applied to 3D animation. In Proceedings of SIGGRAPH '87 (Anaheim FL, July 1987), 35-44.

[18] Laurel, B. Interface agents: metaphors with character. In The Art of Human-Computer Interaction Design, Laurel, B (ed.) Addison-Wesley, Reading MA, 1990.

[19] Lester, J., and Stone, B., Increasing believability in animated pedagogical agents. In Proceedings of 1st International Conference on Autonomous Agents (Marina del Rey CA, February 1997), 16-21.

[20] Lester, J., Converse, S., Kahler, S., Barlow, T., Stone, B., and Bhogal, R. The persona effect: affective impact of

animated pedagogical agents. In Proceedings of CHI '97 (Atlanta GA, March 1997).

[21] Loyall, B. Believable Agents: Building Interactive Personalities. Ph.D. Thesis, Carnegie Mellon University, Pittsburgh PA, May 1997.

[22] Loyall, B., and Bates, J. Hap: A reactive, adaptive architecture for agents. Technical Report CMU-CS-91-147, School of Computer Science, Carnegie Mellon University, Pittsburgh PA, June 1991.

[23] Moon, Y. Can computer personalities be human personalities? In International Journal of Human-Computer Studies 43, 223-239.

[24] Nass, C., Steuer, J., and Tauber, E. Computers are social actors. In Proceedings of CHI '94 (Boston MA, April 1994).

[25] Perlin, K., and Goldberg, A. Improv: A system for scripting interactive actors in virtual worlds. In Computer Graphics 29.

[26] Petta, P., and Trappl, R. On the cognition of synthetic characters. In Cybernetics and Systems '96, Proceedings of the 13th European Meeting on Cybernetics and Systems Research (Vienna, 1996), 1165-1170.

[27] Picard, R. Affective Computing. MIT Press, Boston MA, 1997.

[28] Reilly, W. S. N. Believable Social and Emotional Agents. Ph.D. Thesis, School of Computer Science, Carnegie Mellon University, 1996.

[29] Rickel, J., and Johnson, W. L. Integrating pedagogical capabilities in a virtual environment agent. In Proceedings of 1st International Conference on Autonomous Agents (Marina del Rey CA, February 1997), 30-38.

[30] Rist, T., André, E., and Muller, J. Adding animated presentation agents to the interface. In Proceedings of the International Conference on Intelligent User Interfaces (Orlando FL, January 1997), 21-28.

[31] Rousseau, D., and Hayes-Roth, B. Personality in synthetic agents. Tech. Report KSL-96-21, Knowledge Systems Lab., Stanford University, Stanford CA, July 1997.

[32] Stone, B., and Lester, J. Dynamically sequencing an animated pedagogical agent. In Proceedings of AAAI '96 (Portland OR, August 1996), 424-431.

[33] Thomas, F., and Johnston, O. The Illusion of Life: Disney Animation. Hyperion Books, New York, 1981.

[34] Trappl, R., and Petta, P. (eds.) Creating Personalities for Synthetic Actors. Springer-Verlag, Berlin, 1996.

[35] Zimbardo, P, and Leippe, M. The Psychology of Attitude Change and Social Influence. McGraw-Hill, New York, 1991.