Design and Evaluation of Embodied Conversational Agents: A Proposed Taxonomy

Katherine Isbister
kath@cyborganic.net

Patrick Doyle
pdoyle@cs.stanford.edu
Stanford University
Stanford, CA 94025-9020

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.

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 [3,4,8,12,14,19,20], models of emotion [5,6,18], graphics [2,17] and interface design [11,21], to sociology and psychology [7,15,16,26], and even art, drama, and animation techniques [10, 24]. 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 [1,2,9,23,25]. 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., [13]) 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.

The Taxonomy
Here is our preliminary list of specialties within our field. We hope these will serve as a point of departure for discussion at the workshop. For each, we've included relevant skills, criteria for success, and techniques for evaluation. We listed these working from the 'surface' of the agent, to the beneath-the-surface aspects. Please note: our 'specialties' do not necessarily map to particular established disciplines, but are rather focused on different layers of crafting characters.

1. **Appearance and behavioral believability**

   This specialty is concerned with making a person's visceral reaction to the character a powerful and appealing one. This would include appearance as well as sound and movement—all things that make a character more sensorially appealing. 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.

   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 may use critique by other qualified specialists, to help evaluate success.

2. **Interaction 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).

   Some examples: A researcher who investigates user reactions to an interaction with a character where you type in your responses, but the character talks in response, versus an interaction where both of you communicate in text; a researcher who designs an agent system with two agents that discuss with one another as well as with the end user, to explore the powers of having a pair-to-one interaction; a researcher who examines the social roles that characters play in computer games, towards understanding the range of roles that could apply in other user contexts.

   Skills required: training in researching user needs, desires, and instincts, and designing interactive strategies that map to these. Training in the evaluation and iteration of these strategies.

   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. Also, the creation of broadly applicable design insights from particular projects.

Evaluation techniques: Surveys and rigorous observation of end users in the right target audiences for the interaction at hand; peer critique.

3. **Application contexts**

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: researcher who creates a tutoring agent by carefully researching real tutors in tutoring contexts, and tests the final character in a trial with real students; someone who creates a consumer assistance agent and tests it in a real customer service situation, using benchmark metrics for reducing email volume to that company.

Skills required: domain knowledge about the application area, interaction design and user research skills (or ability to direct such efforts).

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.

4. **Computational techniques**

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.

5. Production

This specialty is often lacking from research efforts, and is not well represented in our projects and papers. It represents the project management skill set that is always a part of industry media creation efforts.

Some examples: 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.

Skills required: organizational skills, project outline skills, people management.

Criteria for success: a completed embodied conversational character, preferably created under sane working conditions within the agreed-upon budget, that accomplishes the primary research objectives set at the outset of the project.

Evaluation techniques: did the methods used produce the success factors mentioned above.

Applying the Taxonomy

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

1. *Use it to clarify primary skills.*
Identify one's area of expertise and deep knowledge in contrast to areas that one knows less about.

2. *Use it to assemble appropriate teams.*
When beginning an embodied conversational character project, think about where one hopes to make a primary contribution or contributions with the project, and compare that to one's own skills. Then, based on this assessment, gather other researchers that have the needed additional skillsets. An alternative may be to get permission to re-use a component from another researcher, to fill a need, 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.

Conclusion

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.

Bibliography

1. E. Andre, Ed. Notes of the IJCAI-97 Symposium on Animated Interface Agents: Making Them Intelligent, Nagoya, Japan, Aug. 1997.

2. N. Badler, "Real-Time Virtual Humans," in Proc. 1997 Pacific Graphics Conf., Seoul, Korea, 1997.

3. J. Bates, A. B. Loyall, and W. S. Reilly, "An Architecture for Action, Emotion, and Social Behavior," Tech. Report CMU-CS-92-142, School of Computer Science, Carnegie Mellon University, Pittsburgh, July 1992.

4. B. Blumberg, "Old Tricks, New Dogs: Ethology and Interactive Creatures." Ph.D. Thesis, Media Lab., Massachusetts Inst. Technol., Cambridge, MA, 1996.

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

6. C. Elliott, J. Lester, and J. Rickel, "Integrating Affective Computing into Animated Tutoring Agents," in Notes of the IJCAI '97 Workshop on Animated Interface Agents: Making Them Intelligent, Nagoya, Japan, Aug. 1997, pp. 113-121.

7. S. Fiske and S. Taylor, Social Cognition. McGraw-Hill: New York, 1991

8. B. Hayes-Roth, L. Brownston, R. Huard, B. Lent, and E. Sincoff, "Directed improvisation," Tech. Report KSL-94-61, Knowledge Systems Lab., Stanford Univ., Stanford, CA, Sept. 1994.

9. B. Hayes-Roth, R. v. Gent, and D. Huber, "Acting in character," in Creating Personalities for Synthetic Actors, R. Trappl and P. Petta, Eds. Springer-Verlag: Berlin, 1997.

10. J. Lasseter, "Principles of traditional animation applied to 3D animation," in Proc. SIGGRAPH '87, Anaheim, FL, July 1987, pp. 35-44.

11. B. Laurel, "Interface Agents: Metaphors with Character," in The Art of Human-Computer Interaction Design, B. Laurel, Ed. Addison-Wesley: Reading, MA, 1990.

12. J. Lester and B. Stone, "Increasing Believability in Animated Pedagogical Agents," in Proc. 1st Int. Conf. on Autonomous Agents, Marina del Rey, CA, Feb. 1997, pp. 16-21.

13. J. Lester, S. Converse, S. Kahler, T. Barlow, B. Stone, and R. Bhogal, "The Persona Effect: Affective Impact of Animated Pedagogical Agents," in Proc. CHI '97 Conf., Atlanta, GA, Mar. 1997.

14. A. B. Loyall and J. Bates, "Hap: A Reactive, Adaptive Architecture for Agents," Tech. Report CMU-CS-91-147, School of Computer Science, Carnegie Mellon University, Pittsburgh, June 1991.

15. Y. Moon, "Can computer personalities be human personalities?" in Int. Journal of Human-Computer Studies, vol. 43, pp. 223-239, 1995.

16. C. Nass, J. Steuer, and E. Tauber, "Computers are Social Actors," in Proc. CHI '94 Conf., Boston, MA, Apr. 1994.

17. K. Perlin and A. Goldberg, “Improv: A System for Scripting Interactive Actors in Virtual Worlds,” Computer Graphics, vol. 29, 1996.

18. R. Picard, Affective Computing. MIT Press: Boston, MA, 1997.

19. W. S. N. Reilly, “Believable Social and Emotional Agents.” Ph.D. Thesis, School of Computer Science, Carnegie Mellon University, 1996.

20. J. Rickel and W. L. Johnson, “Integrating Pedagogical Capabilities in a Virtual Environment Agent,” in Proc. 1st Int. Conf. on Autonomous Agents, Marina del Rey, CA, Feb. 1997, pp. 30-38.

21. T. Rist, E. Andre, and J. Muller, “Adding animated presentation agents to the interface,” in Proc. Int. Conf. on Intelligent User Interfaces, Orlando, FL, Jan. 1997, pp. 21-28.

22. D. Rousseau and B. Hayes-Roth, “Personality in synthetic agents,” Tech. Report KSL-96-21, Knowledge Systems Lab., Stanford Univ., Stanford, CA, July 1997.

23. B. Stone and J. Lester, “Dynamically Sequencing an Animated Pedagogical Agent,” in Proc. 13th Natl. Conf. on Artif. Intell., Portland, OR, Aug. 1996, pp. 424-431.

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

25. R. Trappl and P. Petta, Eds. Creating Personalities for Synthetic Actors. Springer-Verlag: Berlin, 1996.

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