Evaluating Embodied Conversational Agents in Collaborative Virtual Environments

(Michael Gerhard, Fraunhofer ISST, Position Paper for Dagstuhl Seminar 04121)

There are currently no evaluation methods specific to ECAs in CVEs and traditional evaluation methods are limited in their applicability and consequently unlikely to address the full range of aspects now inherent in such systems. We argue that a combination of controlled experimentation, quasi-experiments, review-based evaluation and heuristic expert reviews is needed. To operationalise these traditional evaluation methods the concept of *presence* was deployed, and it was argued that presence as a cognitive variable can be measured and that such a measure can be a key indicator of the usability of ECAs in CVEs. Presence measures can be administered within controlled experiments and quasi-experiments to test certain aspects of the system. Such experiments might turn out particularly useful as a means of selecting between two or more design options and it is argued that issues concerning ECAs in CVEs can be meaningfully evaluated by comparing subjects' experience of presence (Gerhard 2003).

The effect of the deployment of a prototype ECA on subjects' experience of presence was investigated within a controlled experiment. The CyberAxis virtual art gallery (figure 1) was used, consisting of one reception room and three exhibition rooms.

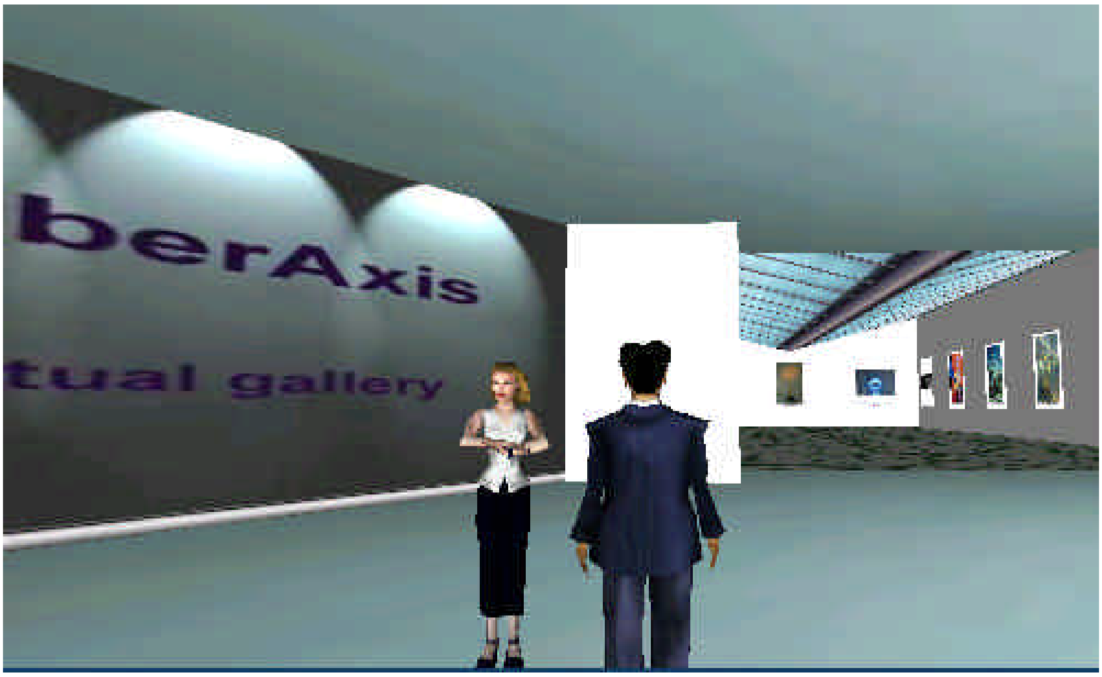

Figure 1

Dagstuhl Seminar Proceedings 04121
Evaluating Embodied Conversational Agents
http://drops.dagstuhl.de/opus/volltexte/2006/460

The *blaxxun Virtual World Platform 5.1* VRML multi-user server was used to make the virtual gallery accessible on a Web server and enable avatar and chat interaction. Part of the *blaxxun* platform is an agent server, which can be interfaced through the *agent.cfg* script file. The agent server, performing event-handling and response selection processes, is responsible for appearance and animation of the agent's avatar. Using the blaxxun agent script it was possible to interface with external applications to extend the functionality of the agent. ALICE bot technology was used to incorporate advanced chat skills and create an embodied conversational agent. The prototype agent was deployed in series of controlled experiments measuring the effects of simulated copresence on subjects' experience of presence. The hypothesised relationship and the combination of all variables involved are illustrated in figure 2.

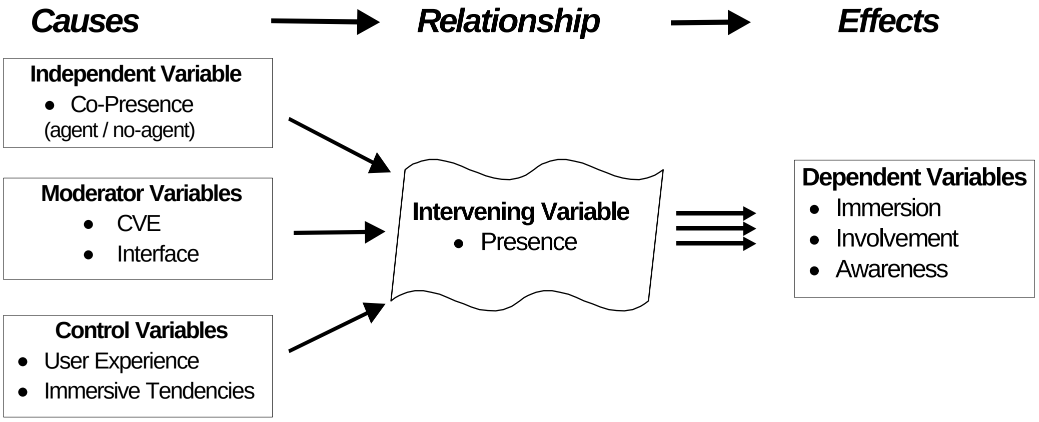

Figure 2

It was predicted that the deployment of an ECA and the experience of presence would be positively related. Results showed that by simulating the copresence of another entity within the experimental environment, the prototype agent did succeed in increasing subjects' experience of presence (Gerhard et al 2004).

Although implementation issues were not the primary concern of this study, the strength and shortcomings of the prototype agent were evaluated as secondary variables within that experiment. Evaluation specifically of natural language systems is a problem that has been studied intensively, with a variety of specific heuristics being applied for such systems (e.g. Hirschman 1998, Polifroni et al 1998). However, a review of the literature could not identify a universally agreed set of criteria for the evaluation of ECAs within collaborative virtual environments. A set of criteria developed for the evaluation of the strengths and shortcomings of the current prototype agent are partly based on Nielsen's (1994) general usability guidelines and partly on a set of heuristics proposed for non-embodied conversational systems (Sanders and Scholtz 1998). Further, the specific set of evaluation criteria identified here is focused on the prototype's abilities that experience and literature (Massaro et al 2000, Oviatt and Adams 2000, Sanders and Scholtz 2000, Nass et al 2000) suggest are important for a successful human-agent dialogue. The criteria identified are separated into those relating to linguistic behaviour and those relating to bodily behaviour.

Linguistic Criteria

L1) Recognition of speech/language input
L2) Generating clear and concise speech/language output
L3) Following human dialogue conventions
L4) Understanding user turns
L5) Understandability of agent turns on a semantic level
L6) Generating output containing relevant information
L7) Generating output containing task related information

These the criteria have been adopted from a wide range of relevant sources. Firstly, as Hirschman (1998) argues, the existence and reliability of speech/language recognition itself is to be assessed at the very basic level of whether a response is triggered at all by users' speech/language inputs (criterion L1). The second criterion (L2) assesses whether the speech/language output generated by the system makes clear and concise use of the English language (cf. Sander and Scholtz 1998). The third criterion (L3) is the ability to follow human conversational mechanisms, which refers for example to the adequacy of turn taking behaviour and the ability to resolve referring expressions, such as pronouns, in the users' turns (Bernsen et al 1997). Understanding of user turns (criterion L4) refers to the agent's ability to utilise pieces of information provided by the user (Polifroni et al 1998).

Criteria L5, L6 and L7 were adopted from a set of metrics proposed by Sanders and Scholtz (2000). The fifth criterion (L5) is the understandability of the agent's turns on a semantic level is, which involves assessing whether output generated by the system *makes sense* to the user in relation to previous statements made by the agent. The sixth criterion (L6) is the frequency of *good* agent turns, which measures the number of turns containing informative feedback, namely information that is intuitively seen as relevant to the environment or to the users' previous inputs. Finally, the seventh criterion (L7) within the linguistic section is the usefulness of agent turns with respect to helping users accomplish their tasks, measured by calculating the frequency of task related system turns.

Criteria Relating to Bodily Behaviour

B1) Generating animated humanoid embodiment

B2) Generating movement within the environment

B3) Generating gestures, including pointing

B4) Generating facial expressions, including eye gaze

B5) Understanding user movements

B6) Understanding gestures, including pointing

B7) Understanding facial expressions, including eye gaze

The existence of an animated humanoid embodiment, i.e. a humanoid face and a full humanoid body, is a pre-requisite of bodily behaviour (Lester et al 2000); it is seen as the first and most basic criterion for ECAs. The second item on this list (criterion B2) refers to the agent's ability to change the orientation and the position of its body within the CVE, which are seen as crucial bodily behaviours (Churchill et al 2000). The agent's awareness and understanding of the movements of avatars representing other users (criterion B5) is also of key importance, since one of the primary advantages of ECAs is their ability to know (and tell users) where things are and how to get around (Rickel and Johnson 2000).

Several studies have shown the importance of the generation and understanding of non-verbal signals within conversations such as facial expressions, including eye gaze, and gestures, including pointing (Ekman 1982, Kendon 1993, McNeill 1992). It therefore had to be assessed whether the agent (criterion B3) as well as the user (criterion B6) were able to talk about objects in the domain of discourse with gesture to indicate mode of interaction, e.g. by saying "turn it like this" and demonstrating how by use of gesture. Further, the prototype was evaluated with respect to its ability to generate (criterion B4) and understand (criterion B7) facial expressions.

References

Bernsen, N., Dybkjaer, H., Dybkjaer, L. (1997) What Should Your Speech System Say?, in *IEEE Computer*, Vol. 30(12) pages 25-31, USA

Churchill, E., Cook, L., Hodgson, P., Prevost, S., Sullivan, J. (2000) "May I Help You?": Designing Embodied Conversational Agent Allies, in *Embodied Conversational Agents*, MIT Press, Cambridge, MA, USA

Ekman, P. (1982) *Emotion in the Human Face*, Cambridge University Press, Cambridge, USA

Gerhard, M. (2003) *A Hybrid Avatar/Agent Model for Educational Collaborative Virtual Environments*, PhD Thesis, Leeds Metropolitan University, Leeds, UK

Gerhard, M., Moore, D., Hobbs D. (2004) (in press) Embodiment and Copresence in Collaborative Interfaces, in *International Journal of Human Computer Studies,* Academic Press, Elsevier Science

Hirschman, L. (1998) Language Leaning Evaluations: Lessons Learned from MUC and ATIS, in *Proceedings of the International Conference on Language Resources and Evaluation,* pages 117-122, ELRA, Paris, France

Kendon, A. (1993) Human Gesture, in *Tools, Language and Intelligence,* Ingold, T., Gibson, K. (eds.), Cambridge University Press, Cambridge, USA

Lester, J., Towns, S., Callaway, C., Voerman, J., Fitzgerald, P. (2000) Deictic and Emotive Communication in Animated Pedagogical Agents, in *Embodied Conversational Agents*, MIT Press, Cambridge, MA, USA

Massaro, D., Cohen, M., Beskow, J., Cole, R. (2000) Developing and Evaluating Conversational Agents, in *Embodied Conversational Agents*, MIT Press, Cambridge, MA, USA

McNeill, D. (1992) *Hand and Mind: What Gestures Reveal about Thought*, Chicago University Press, Chicago, USA

Nass, C., Isbister, K., Lee, E., J. (2000) Truth is Beauty: Researching Embodied Conversational Agents, in *Embodied Conversational Agents*, MIT Press, Cambridge, MA, USA

Nielsen, J. (1994) Heuristic Evaluation, in *Usability Inspection Methods*, Nielsen, J., Mack, R., L. (eds.), pages 25-62, John Wiley & Sons, New York, USA

Oviatt, S., Adams, B. (2000) Designing and Evaluating Conversational Interfaces with Animated Characters, in *Embodied Conversational Agents*, MIT Press, Cambridge, MA, USA

Polifroni, J., Seneff, S., Glass, J., Hazen, T. (1998) Evaluation Methodology for a Telephone-based System, in *Proceedings of the International Conference on Language Resources and Evaluation*, pages 117-122, ELRA, Paris, France

Rickel, J., Johnson, W. (2000) Task-oriented Collaboration with Embodied Agents in Virtual Worlds, in *Embodied Conversational Agents*, MIT Press, Cambridge, MA, USA

Sanders, G., A., Scholtz, J. (1998) Measurement and Evaluation of Embodied Conversational Characters, in *Proceedings of the Workshop on Embodied Conversational Characters (WECC 98),* Tahoe City, California, USA

Sanders, G., A., Scholtz, J. (2000) Measurement and Evaluation of Embodied Conversational Agents, in *Embodied Conversational Agents*, MIT Press, Cambridge, MA, USA