Evaluating ECAs – What and how?

Zs. Ruttkay[§], C. Dormann[¥], H. Noot[§]

§ Center of Mathematics and Computer Science, Amsterdam
{zsofi, han}@cwi.nl
¥ Free University, Dept. of Computer Science, Amsterdam
claire@cs.vu.nl

1.   Introduction

Embodied conversational agents (ECAs) are synthetic characters which can converse with the user (or with other ECAs) by some natural modalities of human-human communication [2]. ECAs have become increasingly popular, as the promising new interface for several traditional computer applications or as the basis for applications in entirely new domains. Much research has been going on to endow ECAs with own behavioral, personality and emotional models, to enable ECAs to act more or less autonomously. The embodiment of the ECAs poses also challenging research tasks, to define which modalities to use to convey some information, how to generate good-quality output in the single modalities (speech, facial-, hand- and body gestures) and how to co-ordinate them. Often ECAs are meant to be involved in conversation with the user, when the maintenance of the discourse is an additional task.

The proliferation of research issues, the diversity of prototypes for different application domains and the multitude of used paradigms and tools makes it difficult to evaluate and compare embodied agents.

In this paper we set out to provide a full checklist to compare ECAs, from four points of view: design, usability, practical usage and user perception. By listing all the factors which contribute to the 'mind' and 'body' aspects of an ECA, we hope to create a common ground to compare ECAs from a technical, design point of view. By identifying usability aspects and methods to measure those, ECAs could be compared from the point of view of usefulness. Thirdly, there are the aspects of how the agent is 'subjectively experienced' by the user. Finally there are aspects of practical applicability like cost etc. Ideally the outcome of dedicated evaluation experiments could serve as 'design guidelines' to define the 'best' ECA for a given purpose.

We are very well aware of the inherent difficulties of finding a common evaluation framework. First of all, the comparison of the capabilities of ECAs to those of humans would require that the multitude of aspects of human-human communication are well described. This is not the case. Actually, ECAs have been used as controllable interlocutors to identify some characteristics of human communication, such as intermodality effects. The other possibility is to rely on usability tests of ECAs. Then the 'what to measure and how' problems arises. While one can come up quickly with aspects like 'ease of use' and 'believability', these concepts are not clearly defined. Moreover, they may have different connotations for experts from different fields (psychology, sociology, ergonomy, and computer science) and for different application domains (e-commerce, banking, tutoring).

In the rest of the paper, in chapter 2 we first provide a check-list of design aspects, concerning 'embodiment', 'mind' and intended usage of an ECA. In chapter 3 we give a structured list of evaluation aspects with clarifying discussion of the concepts used. Finally in chapter 4 we give an overview of works which have been addressing evaluation of ECAs. This list contains only some representative items, because of space limitations. However, we suggest that a complete list of all the relevant works should be done in the similar systematic framework.

It is true for the other chapters as well that they are to be taken as an initial attempt to pin down all aspects of evaluation. We encourage others from the ECA community to contribute, by refining/improving our suggestions. We would like to have a discussion of the issues raised during the workshop, and organize dedicated forums in the future to work out the ideas further. The style of the paper reflects the intention with the content: mostly it has the form of checklists, enumerations of properties and questions.

2.   Design aspects

2.1   The embodiment

2.1.1 Look

Personification Is the body of the ECA to represent a person, or some other living creature (animal), or a non-living object (like Microsoft's paper clip)? In case of a human-like ECA, is it made to be recognized as some well-known real person, or to represent a category of real persons (e.g. by profession, age, gender), or to be an individual new person? In case of a non-human ECA, is it anthropomorphic?

Realism Is the model (meant to be) realistic, or is it artistic, may be exaggerated cartoon-like?

Dimensions The model can be 2D, spruit (2D 'cut-out', which can change orientation), 3D.

Physical details What parts of the body are covered by the model: head, head + neck, torso, full body. What details are given, especially for mouth and hands? Are there wrinkles possible on the face?

2.1.2 Communication modalities

Text or speech An ECA may be designed not to 'tell' anything verbally, or use text bulbs for verbal output, or be able to speak. In the latter case, the quality and the content of the speech is determined by the choice between recorded audio or synthesized speech. Does the speech sound natural, does it express some aspect of the ECA (gender, expertise, personality, emotion)? Is the intonation in accordance with lingual structures and some semantic content (emphasis of new or unusual items)? Is the speech spontaneous (with errors, gap filling sounds, non-speech elements like breath, laughter) or sterile?

Facial display The face can be used to express (exclusively, or in co-ordination with other modalities) several functions. In case of speech output, does the face provide lip-sync, and of what quality? Does it exhibit other phenomena of visual speech, namely providing facial expression for: emphasis, punctuation, state of discourse (turn taking/giving), certain characteristics (size, aesthetics) of objects the ECA refers to verbally, directions, certainty? Does the face (even in the absence of speech) express emotions (which ones), cognitive states (which ones), approval/disapproval? What does the face indicate in its idle state? Are the eyes moving, does the pupil change size? Is the head moving? Are other (may be non-realistic) features used for expressions (hair raising, eyes bulging)? Does a given set of facial expressions get repeated in the same way, or is there some variety? Is superposition and concatenation of facial expressions supported, on what basis? Can the face change color (redden, turn pale), and reflection (sweat)? Are the facial expressions meant to be realistic, may be characteristic of a given real person, or of some group (by culture, by profession), or generic? Are the facial expressions designed as cartoon-like?

Hands Are hands used in coordination with speech, to structure and punctuate speech (beat, gestures for enumeration, contrast, change of topic, dialog turns)? Are hands used to point, if so, to what? Are emblems used (which ones), metaphors to indicate characteristics (like form, motion and temporal aspects)? Are hands used (alone, or together with body and/or face) to indicate emotional and cognitive states (which ones)? Are hands used to demonstrate certain specific actions, to manipulate objects? What varieties of a hand gesture can be used, concerning target of the gesture and manner of the motion?

Body Are body postures used in coordination with speech, to indicate change of topic, dialog turns? Does the torso and the body move in accordance with hand gestures? Is the body used to express physical, emotional or cognitive states (which ones)? What about idles state? Can the character change location, in what way (sliding, walking, running) and in what space? What other movements can the body do? Is the body movement repetitive or varying? Is it typical of a real person, or a group (profession)?

2.2   The mind

Humans communicate in a subtle way: they express the very same content differently, depending on their own mental and physical state, the characteristics of and relationship to the person they are addressing, more or less strict cultural and social regulations, the physical (noise, visibility) and social (public versus private space) characteristics of the location. In case of an ECA, it is not enough to make sure by the design of embodiment that the ECA can realize such subtleties, but it is necessary to provide mechanisms to decide about the choice of modalities and expressions appropriate in the given situation. The knowledge and mechanisms to do so are referred to as (part of) the 'mind' of the ECA.

Personality, intention Is the ECA designed to have a certain personality? What personality model is used? In what aspects of the communication is the personality manifested (ideally, in all)? Is there a goal which the ECA is to achieve (a 'salesmen' ECA being more tolerant and friendly with the interlocutor in order to sell, versus 'tutor' ECA)? Is time critical in achieving the goal, is it taken into account?

Emotion model What emotional states can the ECA get into? Are the possible emotional states exclusive categories, or mixtures? Are triggers of emotions learnt by experience? How is the change of emotion in time modeled: what triggers emotions, how do emotions die out?

User model Does the ECA maintain a model of some aspects of the user, like: expertise in the domain, age, gender, ethnicity, cultural and socio-economic background, and more importantly personality traits (such as cognitive style, locus of control, anxiety, etc.)
How is this acquired: by asking for the user profile, by learning? In what way does the model of the user influence the communication of the ECA?

Model of operating environment Are physical characteristics of the environment of the user (hardware configuration, noise) and of the 'world the ECA is in' taken into account?

2.3   Control and interactivity

Reaction mode An ECA may be controlled off-line or on-line. In the first case, the 'content' to be presented by the ECA, and eventually other information on the presentation, are to be given in advance. In what form and detail should the relevant information be given? How long does it take to specify a typical input, and then the ECA to react? What – specific or common (may be standard) – format is to be used to specify input and control? In case of reactive ECAs, is the behavior real-time, or 'acceptably prompt'?

Control Who controls the ECA: some application directly (presentation ECA), the user directly (avatar in virtual forums), or both (educational ECA)? In the latter case, is there an explicit discourse model used? Do aspects like user model, emotional state influence discourse strategy in a static or dynamic way? Is the ECA prepared to recover from erroneous input (content, timing), react to lack of input (after some time)? Is it indicated clearly who's turn it is?

Input modalities of the user Though perception of the interlocutor plays an important role in human-human communication, current ECA design has been concentrating on its presentational aspects, probably because of the technological bottleneck in perception. However, for reactive ECAs and for a symmetrical role in the interaction, it would be beneficial to endow ECAs with perception and sensing capabilities. So it should be a design concern to define how and what is perceived of the user.

2.4   Application context

The application context decides, by and large, what characteristics an ECA should have. We distinguish: presentation ECAs, information ECAs, educational ECAs, sales ECAs, entertainment ECAs and research ECAs to learn about (multimodal) communication.

3. Evaluation

In this chapter we will identify some aspects relevant for evaluation, and discuss what methods could be applied, as well as other evaluation concerns (like dependency of parameters). As already discussed in the introduction, human-human communication and hence human-ECA communication is extremely complex, many parameters are involved, several of which are not clearly understood or maybe even unknown. Moreover, it is very difficult to separate the effect of the application and the effect of the ECA. On the other hand, in case of different applications different aspects of the ECA may be relevant. Finally, different user groups have different expectations from and reactions to an ECA.

We list empirical data collection and evaluation methods to test usability of ECAs and to evaluate how they are perceived. The web-site [12] provides resources on evaluation, in general.

3.1 Methodology of evaluation

3.1.1 What to compare an ECA to?

An ECA can be evaluated as a means for using a computer system, in comparison with traditional systems for the same task. The main question is if the ECA really provides added value, in terms of usefulness, or the scope of potential users. If an ECA is compared to other ECAs, then the question is which is the better/best ECA for the given situation? Because of the many parameters and evaluation criteria, such comparisons address only a subset of the configuration of design parameters.

3.1.2 Testing by what users?

When performing usability tests, the group of subjects should be selected carefully. Some aspects of evaluating ECAs (e.g. trust, engagement) are directly related to the personality of the subject. Others may be indirectly influenced by user aspects like expertise with computers and education. The experimental subjects should be representative of the intended users. (Which is often not the case, when subjects are recruited from university students and staff around.) Users' characteristics to take into consideration include: age, gender, ethnicity, cultural and socio-economic background, personality traits and computer experience. ECAs are novel phenomena.

One should take into account the 'novelty' effect. Also, a critical usage time (which depends on the application) is needed to be able to conduct experiments with realistic usage situations.

Finally, one should consider if the subjective perception of the user corresponds to reality (e.g. an ECA is perceived as trustworthy, though it was not designed to have such a characteristic).

3.1.3 Empirical data collection

Observation In observational evaluation end-users are studied undertaking real-task scenarios either in the workplace or in the laboratory with the observers making notes about interesting behavior or recording their performance for example on video.

Experiment In experimental evaluation users are involved as subjects in a controlled way, in order to provide empirical support for a particular claim or hypothesis concerning some aspect of the ECA. Systematic evaluation of the design criteria can take place in experiments where different ECA designs are tested, by concentrating on single parameters. Such experiments are also meant to shed light on characteristics of human-human communication.

Benchmarks and comparative tests These are standardized forms of experimental procedure consisting of monitoring user's performance on carefully constructed standard tests. It would be very useful to compare ECA applications in a structured way on one (or a few) dimensions. It is still a challenge to define 'benchmark scenarios' to test different aspects of ECAS. Scenarios for ECAs made for different application types should be given. Is this possible at all?

Survey and online survey On-line survey consists most frequently of closed questions (yes/no, multiple choice, short answer) and it is good for wish lists, attitudes, experiences; not for actual behaviors. The advantages are that users may be located anywhere and that a large number of responses can be gathered.

Questionnaire Subjects answer in writing a fixed series of simple, short questions, which can be closed or open-ended. In the first case, the possible answers can be a scale of some aspect, or of several aspects, or a single or multiple-choice list.

Interview Similar to questionnaire, except that the researcher puts down answers to the more elaborate questions. The order of questions or the exact wording can be adapted to the subject or the situation. An interview is usually carried out with smaller numbers of participant (e.g. 15) than a questionnaire. This method gives the richest data and following up on questions is possible. It is good for example for collecting answers on user's attitudes and experience in using ECAs.

Focus group This method consists of a discussion in a small group, moderated by a trained facilitator. It is useful for gathering user impressions and attitudes toward ECAs. The discussion is influenced by group dynamics however.

Usage data It is a list of quantitative characteristics of interaction of the user, produced e.g. on the basis of logged user's action. It often complements other evaluation results.

3.1.4 Measurement techniques

There are different methods to evaluate answers to closed questions. The possible answers provide rating of some aspects, either in a quantitative or in a qualitative way, by offering labels to choose from. If one has to rank with respect to two opposite alternatives, then multi-point rating scale or Likert-scale can be used. Different aspects of an ECA can be measured by 'placing' the ECA in a multi-dimensional space of aspects, using semantic differential (SD) scale.

3.2 Evaluation of usability of ECAs

In this and the next chapter we discuss evaluation of ECAs from the points of view of how easy the user can work with the ECA (chapter 3.2) and how it is experienced by him (chapter 3.3) We will present the major evaluation dimensions, some of which can be decomposed into further, lower-level criteria. E.g. believability is possibly a precondition for trust. Moreover, some criteria may not be completely independent of each other: e.g. believability and engagement are likely to be related in many application domains.

3.2.1 Learnability, ease of use

Definition The ease/difficulty to become familiar with the ECA, from the point of view of communication.
Discussion Is the interaction with the ECA easy to learn? Does it seem to happen in a natural way? Do the (communicational and 'mind') limitations of the ECA become clear?

3.2.2 Efficiency

Definition The degree to which the ECA enables the task to be completed in a timely, effective and economical fashion.
Discussion How fast is the task carried out with the help of the ECA? Is the dialogue with the agent adequate? Is the system flexible enough? Efficiency can influence the acceptance of agents by users. A common measure of efficiency is the time spent on the task, in comparison with using a traditional system for the same task. The pitfall of this measure is that the user may spend extra time with the ECA because he likes it or finds it interesting. Another measure of efficiency is the quality of the task completion.

3.2.3 Errors

Definition The relative amount and type of errors which occur while interacting with the ECA.
Discussion what kinds of errors occur: misunderstanding of the ECA or of the user, incorrect input, miscomprehension in the dialogue (on both sides), time management confusion, deadlock situation? How much time of the total interaction is spent on error situations?

3.3 Evaluation of user perception of ECAs

3.3.1 Helpfulness

Definition Helpfulness is the perception that the ECA communicates in a cooperative way to assist the user in achieving his goals and resolving difficulties.
Discussion Does the ECA provide assistance for example with helpful hints when user appears to experience problems? Does the user 'feel in control', or does he often get lost, not knowing what to do next?

3.3.2 User's satisfaction

Definition Though one of the most often measured aspects, user satisfaction is a compound concept, incorporating aspects such as likeability and affect (the respondents emotions towards the ECA), attractiveness, personality, etc.
Discussion Is the user pleased with using the ECA? Would he prefer to use it in the future, in place of traditional application?

3.3.3 Believability

Definition The ECA is believable if it acts according to the expectations of the user.
Discussion The user judges the ECA based on its look and communicational behavior. These should be consistent at each moment and along time. Believability in general is not equal to 'taking the ECA as of a real human'. This is only the case if the look (realistic, 3D) and communication capabilities both are aimed at the level of a human in a similar role. This is not the case with today's ECAs, which underlines the importance of believability as an evaluation criterion. A realistic 3D head with robot-like speech is expected to be less believable than a robot face with the same speech. It is a complicated and very interesting question when, and to what extent is the user expecting human-like behavior, even if the embodiment makes it clear that the ECA is not meant to be mistaken for a real human. On the other hand, it has been shown that the closer the look is to realism, the more critical the user is in judging it believable.

3.3.4 Trust

Definition Trust is the individual's beliefs about the extent to which a communicating partner behaves in a way that is benevolent, competent, honest or predictable in a situation.
Discussion The definition of trust reflects the multiple study perspectives in different fields. Views of trust as the willingness of the user to make himself vulnerabile, the reliability, familiarity, and solidarity of the ECA can all be applied to ECA applications. Trust in an ECA is primarily based on the content that the ECA communicates to the user, which is determined by the content and services of the application behind the ECA. However, design factors like personality, look, technical quality of communication may contribute to trust.

3.3.5 Engagement.

Definition Engagement (involvement, appeal) indicates how much the user is attracted by the ECA, how close or distant he feels to the ECA [5].
Discussion Both the relevance of the services of the ECA and its design aspects (aesthetics of body, gesturing and speech) and its personality have an effect on engagement. Engagement is related to believability.

3.4 Practical applicability

A final dimension of evaluation is the practical applicability: the conditions and resources need to (re-) use the ECA. Relevant aspects are needed hw/sw resources and their costs, the robustness of the ECA, the user target group and the adaptability/configurability of the ECA.

4. Comparison of evaluations

The table below is meant to serve as a crisp and systematic framework to present works on evaluation. We have added some illustrative items from the substantial (but often hidden as a single chapter descriptions of ECA applications) work on evaluating ECAs.

Ref	Changed parameter	Evaluated parameter	Method of data collection/eval.	Subjects	Application	Findings
Nass et als. [12]	personality: introvert/extrovert (by posture) ethnicity (by look) (in)consistent verbal/nonverbal	Trust liking	questionnaire	40 Korean students 40 students (extrovert/introvert)	application independent ('item selection', arguing)	more trust in extrovert and identical ethnicity ECAs
Cassell et al. [3]	envelope emotional feedback	ease of use efficiency lifelikeness	Survey + performance data analysis	12 novice comp. users, native English speaker	Information provider about the solar system	Envelope is more important than emotional feedback
Colburn [4]	Eye-gaze	Eye response by user Turn taking	Analysis of Eye pattern of user	20, CS staff	Casual chat	For Lester avatar with eye-gaze, users respond with eye-gaze
Lester et als. [8]	Pedagogical agent with different modalities	Effectivity Likeing Entertainment etc. (18 aspects)	Performance test Data analysis	100 secondary-school students/ novice-expert student	plants	Lifelike agent has positive effect on learning: -performance -experience depending on expertise of student
Koda et als. [7]	smiley/dog/ realistic&cartoon man/woman/ no face	Involvement likability	usage data registration + online questionnaire via Internet	157 out of 1000+ users, mostly men	poker game	face is engaging, likable and comfortable all faces were attributed with intelligence, realistic ones the most
Mc Breen [10]	3d woman/man formal/informal appl. domain	aspects of liking trust	questionnaires	36 subjects	banking/cinema/ fligh	Trust less in case of banking appl. dress requires according to appl.
Mc Breen [9]	Video/talking head/still with moving lips/voice	liking	Questionnaires + focus group		Textile e-retail	Video liked best, talking head least! voice only was liked
Bick-more et al. [1]	smalltalk	Trust liking	questionnaires + analysis of behavior	18 students	House sail	smalltalk induces trust with extrovert subjects
Walker et als. [13]	no face/ neutral face/stern face	liking effectivity	questionnaire	49 adults from CS research environment	filling in the questionnaire	voice only least liked/ inefficient; neutral face liked most, stern face was efficient

References

1. Bickmore, T., Cassell, J. (2001) Relational agents: a model and implementation of building user trust. Proceeding of CHI'2001, Seattle
2. Cassell, J. Sullivan, J., Prevost, S., Churchill, E. (Eds.) (2000) Embodied conversational agents. Cambridge, MA: MIT Press.
3. Cassell, J. and Thórisson, K.R. (1999). The power of a nod and a glance: envelope vs. emotional feedback in animated conversational agents, Applied Artificial Intelligence 13: 519-538.
4. Colbrun, R. A., Cohen, M., Drucker, S. (2000). M. The role of eye gaze in avatar mediated conversational interfaces, MSR-TR-2000-81, Microsoft
5. Hoorn, J. (2001) Personification: merger between metaphor and fictional character. Submitted for publication. http://www.cs.vu.nl/~jhoorn.
6. Ibsister, K. Nakanishi, H., Ishida, T., Nass, C. (2000) Helper agent: Designing an assistant for human-human interaction in a virtual meeting space. Proceeding of CHI'2000, pp. 57-64.
7. Koda, T., Maes. P. (1996) Agents with faces: The effects of personification of agents. Proc. of HCI'96.
8. Lester, J., Converse, S., Kahler, S., Barlow, S., Stone, B., Bhogal, R. (1997) The Persona effect: affective impact of animated pedagogical agents, Proc. Of CHI'97, pp. 359-366.
9. McBreen, H.M., Shade, P. Jack, M.A., Wyard, P.J.(2000) Experimental Assessment of the Effectiveness of Synthetic Personae for Multi-Modal E-Retail Applications, Proc. Fourth International Conference on Autonomous Agents, pp.39-45.
10. McBreen, H., Anderson, J., Jack, M. (2001) Proc. Workshop on Multimodal Communication and Context in Embodied Agents, Autonomous Agent 2001, pp 83-87.
11. Nass C., Isbister K., Lee E. (2001) Truth is beauty: researching embodied conversational   agents. MIT Press   Cambridge, MA, USA
12. Usability Testing Resources: http://jthom.best.vwh.net/usability/usable.htm
13. Walker, J. H., Sproull, L., Subramani, R. (1994) Using a human face in an interface. In Human Factors in Computing Systems, pp. 85-91.