Chapter 2

EMBODIED CONVERSATIONAL AGENTS ON A COMMON GROUND

A Framework for Design and Evaluation

Zsófia Ruttkay, Claire Dormann, and Han Noot

> *The one who seeks truth is a scientist. The one who wishes to realize the free flow of his subjective thought is a writer. But what can be done if one needs a way between these two possibilities?*
>
> —Robert Musil

Abstract One would like to rely on design guidelines for embodied conversational agents (ECAs), grounded on evaluation studies. How to define the physical and mental characteristics of an ECA, optimal for an envisioned application? What will be the added value of using an ECA? Although there have been studies addressing such issues, we are still far from getting a complete picture. This is not only due to the still relatively little experience with applications of ECAs, but also to the diversity in terms and experimental settings used. The lack of a common, established framework makes it difficult to compare ECAs, interpret evaluation results and judge their scope and relevance. In this chapter we propose a common taxonomy of the relevant design and evaluation aspects of ECAs. We refer to recent works to elicit evaluation concepts and discuss measurement issues.

Keywords: Embodied conversational agents, design, evaluation framework, methodology.

Z. Ruttkay and C. Pelachaud (eds.), From Brows to Trust, 27–66.

1. Introduction

In this chapter we set out to provide a framework to evaluate and compare ECAs. We undertake this task with the following objectives in mind:

- We wish to provide a framework to categorise the extensive literature on ECA design and evaluation and hence to help us in interpreting and understanding the findings reported.
- We encourage the ECA community to start agreeing upon a common set of concepts used to report on ECA research. This will make comparison of results (much) more meaningful than it is now.
- Hopefully, the end-result of the use of a common framework by the ECA community will be the emergence of design rules for ECAs stating what properties an ECA should have in order to fulfil certain functions.

We are well aware of how challenging and ambitious such a task is. One might ask if it is a realistic and timely task at all. Yet we are convinced that a common evaluation framework will facilitate the judgement and proliferation of empirical results and theoretical guidelines, as well as help to identify fundamental research to be done on specific characteristics to such an extent that it is certainly worthwhile to start developing such a framework now. On the other hand, our framework put forward here will probably need some refinement and readjustment, as more academic results will be available on human-human communication and more empirical evidence will be collected on using ECAs in all kinds of application domains

When proposing a common framework, we do rely on the work done so far. Namely, we have done our best to locate all recent works addressing evaluation of ECAs. Dehn at al. (2000) give a critical summary of works done earlier. We have used the relevant studies from the ECA literature to elicit concepts, to point out controversial issues and draw attention to methodological problems. However, the references are meant to be illuminative, and not to give a complete list of all occurrences of certain evaluation issues.

An ECA can be considered as a novel user interface. We have examined if we could profit from established user interface evaluation methods in HCI. However, in the case of ECAs it requires extra effort and attention to separate the cumulative effect of the underlying application, of the mental and of the embodiment aspects of the ECA.

In the next section of this chapter, we discuss ECAs from a design perspective. First we give a general description of the software environment in which we envisage an ECA to operate and define the concept of an ECA by delineating it from the other software components it interacts with. Then review all the properties of ECAs which may be relevant for comparison of existing ECAs and specification of new ones with certain expected functions. In section 3 we turn to the methodological aspects of evaluation of an ECA, discussing critical issues as setting base-line for evaluation, the types of evaluation studies and design guidelines abstracted and the problematic of defining evaluation concepts. We outline the characteristics of tests subjects which may influence the evaluation, and methods available to collect and evaluate empirical data. Section 4 is devoted to the definition and discussion of concepts relevant to evaluate ECAs. In the concluding section we give a summary of the long-term potentials of our proposed framework, and make some concrete recommendations on ECA evaluation.

In the rest of this section we give the motivations for our endeavour.

1.1 Motivations and Problems

The evaluation of the capabilities of ECAs in the light of those of humans would require that the multitude of aspects of human-human communication have been described in a normative way and with the granularity matching the design parameters of ECAs. This is not the case. Unfortunately, there are not enough sources from the fields of socio-psychology, sociology, cultural anthropology and psycho-linguistics to rely upon for a complete description of, for instance, what a tutor should look like, how he should talk and gesture, given an application domain and a target group. Actually, the introduction of ECAs has motivated research in human-human communication, by posing new, succinctly formulated questions, some of which could be answered only by using ECAs as controllable mediums that exhibit the effects to be tested (see Chapter 7 by Krahmer et al. in this book). Moreover, it has to be justified if it is a correct objective to try to mimic human behaviour when creating ECAs. The technology does allow the creation of non-human, non-realistic creatures, but the problem of devising the 'right' communicational skills for such creatures and evaluating their merit is no less challenging.

One could rely on usability tests with the ECAs developed so far. Then the 'what to measure and how' problem arises. While one can come up quickly with aspects like 'ease of use' and 'believability' of the ECA as desired objectives, these concepts are not clearly defined. Moreover, they may have different connotations for experts from different fields as psychology, sociology, ergonomy, and computer science. These concepts are likely to have different interpretations depending on the application domain, such as e-commerce, banking or tutoring. One cannot be sure if the similar concepts reported in different studies were used with the same meaning. Moreover, the diversity in the settings for empirical data collection and in the evaluation methods used makes one uncertain if a reported conclusion is sound and general enough to be taken as a design guideline.

Finally, there has been relatively little done on ECA evaluation, and with a series of different objectives. Some researchers, interested in the potentials of applying ECAs in a specific domain, or endowing an ECA with mechanism to exhibit some specific characteristics, collected empirical data to test how people react to the ECA with the new feature. These reports are typically found as one of the last sections of a paper, and often account on experiments done with one or two dozens of computer science students as test subjects. Since a few years ago one can read more extensive works dedicated per se to evaluation of ECAs used in operating environments (Moundridou and Virvou (2002); Buisine et al. (2003); Bickmore and Cassell (2003); Lester et al. (1997); Cassell and Vilhjálmsson (1999); Höök et al. (2000); Isbister and Hayes-Roth (1998); McBreen et al. (2000); McBreen et al. (2001); Mori et al. (2003)) or to figure out how basic design parameters for an ECA influence the users impression (Barker (2003); Cassell and Bickmore (2000); Cowell and Stanney (2003); Isbister and Nass (2000); King and Ohya (1996); Koda and Maes (1999); Nass et al. (2000); Nass and Lee (2000); Sproull et al. (1996)). Only recently, some researchers of ECAs have addressed evaluations dimensions and methodologies as such (Sanders and Scholtz (2000); Isbister and Doyle (2002); Chapter 9 by Catrambone et at. in this book).

2. ECAs from a Design Perspective

The user will react to an ECA based on both *what* it communicates, and *how*. To differentiate between the matters of producing syntactically correct output signals by using one or more modalities to present some message and the matters of deciding what to express, in the literature the *body* and *mind* distinction has been used. The mind aspect has been associated by Pelachaud et al. (2002) with *reasoning* and the AI techniques used to implement reasoning. In our discussion, we keep the *body aspects* but replace the mind with the *mental aspects* concept. We wish to have a broader category encompassing also phenomena like personality, which are static and do not necessarily involve the kind of intelligence and reasoning associated with the mind. Moreover, in our design-oriented discussion we are not concerned with the underlying *mechanisms* of triggering the communicational behaviour of the ECA, but only with the *effect* of it. Barker (2003) claims the 'illusion of life' can be achieved without any cognitive processing mechanism, by carefully designing the embodiment of the ECA.

We set our focus for ECA evaluation by concentrating on *design aspects*: what is the effect of certain characteristics of an ECA? The *embodiment design parameters* define the *look* of the body (static characteristics like gender, race, cartoon or realistic design) and the *capabilities of the communication modalities* (dynamic characteristics of facial and body gesturing). The *mental design parameters* are responsible for *conversational*, *personality* and *social role* characteristics. These parameters will have an effect on how things get presented for the user. In order to delineate the topics discussed in this chapter, we describe a conceptual architecture of the ECA and the assumed software environment in which it operates. Note that we only consider the ECA in its role as communicating to the user, about the communication channel from user to ECA analogous remarks can be made.

In Figure 2.1 we give an overview of the aspects of an ECA to be dealt with. The following steps are relevant in determining the behaviour of the ECA:

1 At the basis there is an *application* which produces information. This output may be in a textual form, close to one used in human-human communication (e.g., news items collected from the web), or data in a coded form (e.g., time-table items, numerical values of measurements, images, video etc.).

2 The agent translates the content provided by the application into a form which can be used for presentation for the user. This translation is done by using (one or more) *application interface* modules; resulting in a content the ECA can deal with further. One such form is text marked up with meaning tags, expressing different meta-information on the content.

3 The *agent*, with the use of its mental capabilities, decides about when and how the content is to be presented. Two types of task are essential:

   - coordinating the communication between the user and the application.
   - presenting the information provided by the application interface.

In its simplest form the agent just transmits information between the two (while maybe changing some formats); in a complex form the agent is truly autonomous and proactive. In that case it may for instance monitor the user's activity to determine when to get active.

4 The ECA performs the presentation, by using the possibilities of its *embodiment*. Besides the dynamical characteristics of the verbal and nonverbal presentation capabilities (e.g., facial expressions, speech, gestures), the static characteristics (the look) will also contribute to the impression it makes on the user.

2.1 The Embodiment

We use the term *embodiment* in a broad sense, for all low-level aspects which contribute to the *physical appearance* of the character, namely: body design and rendering, voice, head, face, hand gestures and body postures, the quality of the corresponding motions. Each of these aspects may have an effect on the perception of mental aspects of the ECA, or directly on the performance effect achieved by the ECA.

2.1.1 Look

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

The majority of ECAs are designed with a human look, with attributes suggesting a professional role like medical consultant, sales assistant, newsreader, or representing the user in virtual worlds or chat forums. There is cautiousness with applying and evaluating non-human living characters; we know of dogs (Isla and Blumberg (2002); Isbister et al. (2000); Koda et al. (1996)). The reason for this can be in the hidden assumption that "the more human-like the ECA is, the better". This assumption is not justified, in this generality. People attribute more intelligence and trust to human-like ECAs (King and Ohya (1996)), but a (well-designed) non-human character may be more appealing and entertaining. Moreover, in one case the dog appearance was chosen (Isbister et al. (2000)) to avoid that users assume and expect highly intelligent mental capabilities from the ECA. As of objects, we have Microsofts paper clip. (Unfortunately, we cannot refer to studies on its popular-

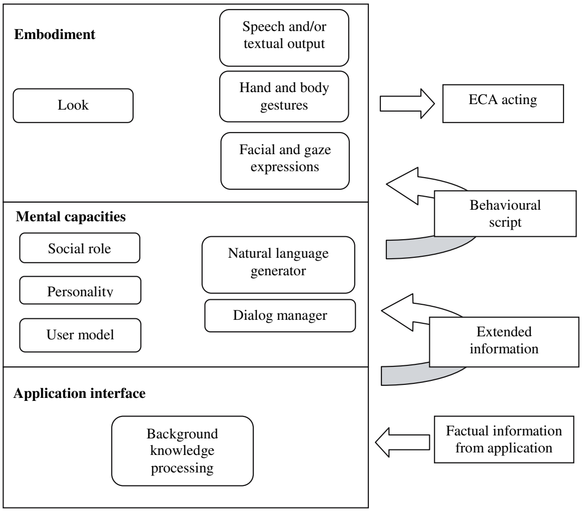

Figure 2.1. The conceptual framework for design aspects of an ECA.

ity, and wonder if the embodiment has been evaluated in any stage of its design). It is a challenge to find out which objects are appropriate candidates as ECA embodiment, both from technical points of view (they should have some face, some means of gesturing, some coding of human-like expressions) and of user reactions.

Physical details What parts of the body are present in the model: head, head plus neck, torso, full body.

Faces have been extensively used, due to the attractive power of the human face. A common application is a talking head, to enhance the intelligibility of speech (see Massaro (1998)). The application context may make it clear if hands (e.g., used for pointing) or the full body (e.g., to change location) are an absolute necessity. In general, it is not true that "the more of the body used the better" is a good design principle. There are some experiments confirming that users spend most of the time looking at the ECA's face (Witkowski et al. (2001)).

Realism Is the model (meant to be) realistic, or is it artistic, may be exaggerated cartoon-like? Is the level of realism the same, or is there a realistic face on a cartoon-like body?

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

General deformability What features can be moved and deformed? Does the model provide seamless joints, wrinkles (on the face)?

2.1.2 Communication Modalities

Language In what natural language (NL) does the ECA communicate? How rich is his language repertoire? How many different things can he express, in what verbal variations? How is the language output generated: selected from pre-defined samples or patterns, or generated on the fly by some NL generator? Is the language usage (words, grammatical structures) designed to reflect some mental characteristics of the ECA? Does the language usage of the ECA change according to some dynamical parameters of the user and/or of the ECA?

The language usage, though in itself a complex issue and usually taken for granted by using some existing NL module, cannot be discarded when evaluating ECAs. First of all, the language used (e.g., English) may imply some cultural connotations for the ECA. According to Isbister et al. (2000), English as the language of communication may be a bias for the Japanese users when interacting with a conversational mediator ECA which was designed to bridge cultural and communicational gaps between Japanese and American users. It is a subtle but important point to remember, even if we tend to believe in such a statement as “the language of the Internet is English”. Further, according to Prendinger and Ishizuka (2002) language is powerful in conveying personality and social role aspects of the ECA.

Textual or verbal output An ECA may be designed not to ‘tell’ anything, just to be present and communicate without words. But in most of the cases, an ECA is endowed with a separate text window or body-related text bulbs for verbal output, or is able to speak. In the latter case, are the utterances produced as pre-recorded audio, or generated by some text to speech (TTS) engine? How understandable is the (synthesized) speech? Does the speech sound natural? Is it in accordance with the static embodiment (gender, age) and mental characteristics (personality, social role) of the ECA? What can be expressed by meta-speech characteristics (intonation, speech rate, etc), for instance punctuation, emphasis, emotions, certainty? Is the speech spontaneous (with errors, gap filling sounds, non-speech elements like breath, laughter) or ‘perfectly sterile’?

With the development of the quality and accessibility of synthesized speech, talking is becoming a common modality of an ECA. All the same, one can still find examples of text bulb usage or even textual communication in a separate window, as examples of output redundant to speech. The pros (increased intelligibility) and cons (extra mental load) of using both speech and textual output, as well as the ideal design for the content, form and function (e.g., numerical data are shown only, in form of tables) are to be investigated.

As to the importance of tuning meta-speech characteristics of synthesized speech, experiments by Nass and Lee (2000) have shown that users do recognize personality characteristics in synthesized speech, and presenting personality in speech modality alone is sufficient to induce a different personality image of the ECA, and thus a different reaction by the user.

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? Can it exhibit other phenomena of visual speech, namely providing facial expression for: emphasis, punctuation, regulation of discourse, conversational feedback, certain characteristics of objects the ECA refers to verbally? Can the face express emotions (which ones), cognitive states (which ones)? What does the face indicate in its idle state (what expression, blinking and head motion)? Do the eyes move and the pupils change size? Does the head move? Are other, maybe non-realistic features (like hair rising, eyes bulging) used for expressions? 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 colour (redden, turn pale)? 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?

The effect of speech punctuation by facial expression has been investigated, recently in more depth and for different cultures too (see Chapter 7 by Krahmer et al. in this book). It has been shown that a subtle and static difference in the basic expression results in difference in the effectivity of task performance and in subjective impression of the task (Sproull (1996)). The facial display has been shown to be successful in expressing friendly-unfriendly personality aspects (Prendinger and Ishizuka 2003). Gaze has been shown to be relevant for dialogue regulations, and expressing personality (Krahmer et al. (2003)).

Hands Are hands used in coordination with speech, to structure and punctuate speech (beat, gestures for enumeration, contrast, change of topic, dialogue turns)? Are hands used to point, if so, to what, and in what way (precision)? 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?

Body Are body postures used in coordination with speech, to indicate change of topic, dialogue turns? Does the body move in accordance with hand gestures? Is the body used to express physical, emotional or cognitive states (which ones)? How about the idle state? Can the character change location, in what way (sliding, walking, running) and in what space? What other movements can the body perform? Are body movements typical of a real person, or a group (e.g., of the same profession)?

Modality coordination and motion generation How are the different modalities used? What aspects of the ECA (such as personality, social role) are reflected in the modality usage and motion characteristics of the gestures? Are there stills used, or animations? What are the motion parameters of animations? The simplest case is when some (single or fixed set of) modalities with given animation are used to express some meaning. In a more sophisticated scenario the selection and fine-tuning of the gestures is done dynamically, according to the characteristics of the situation. How are the problems of blending and concatenating gestures and channel allocation conflicts solved?

2.2 The Mental Aspects

Humans use the body and the voice to express different aspects of a piece of factual content, according to a given situation. The knowledge and mechanisms of an ECA to enhance factual information with meta-aspects like emotions or certainty are referred to as (part of) the mental capabilities of the ECA.

2.2.1 Personality Is the ECA designed to have a certain personality? What personality model is used? In what aspects of the embodiment (speech characteristics, gestures, postures, design of look) and other mental aspects (wording and structure of language used, dialogue management) is the personality manifested?

There exist established personality models in psychology. Probably the one most used in the field of ECA evaluation is the five factors model (see McCrae and John (1992)). The factors are agreeable, extroverted, neurotic, conscientious and open. There exists abundant evidence based on empirical evaluation that there are strong interactions between user personality, perceived ECA personality, and subjective evaluation dimensions discussed in 4.2. (see Dryer (1999); Nass and Lee (2000); Cassell and Bickmore (2000)). So ECA designers should take the ECA's personality in this technical sense serious, and perform the required evaluations.

2.2.2 Social Role The social role of an ECA roughly corresponds to some professional category like teacher, salesperson, and clerk. However, these roles can be further refined, e.g., in the case of teacher to 'expert providing professional feedback' or 'educator providing motivational feedback'; or 'ally for the user' versus 'examiner of the user'. In the above sense, it is important to design the intended social role of the ECA, and reflect it in its embodiment and mental aspects.

Ideally, the manifestation of the social role in behavioural and presentation parameters should be evaluated (and, maybe also, designed) with reference to sociological and anthropological investigation. Isbister et al. (1998) give an example of such an evaluation when they analyse the behaviour of an agent playing the role of a bartender in a chat application. Prendinger and Ishizuka (2002) report on the perceived social role effect (power relationship to user) of ECAs.

The ECA technology allows cases without parallels in real life. For instance, in a real shop the user is communicating with a single salesperson who acts according to some mixture of his own interests of selling certain items and of the interest of the user. The two interests can be manifested in two ECAs, confronting positive and negative aspects of products, as shown by André and Rist (2000).

2.2.3 Emotions What emotional states can the ECA get into? Are the possible emotional states exclusive categories, or mixtures? Is there some emotional model used, also for changes in emotions? Is it verified that the emotions the ECA is claimed to have are indeed perceived as such by users? When the ECA may be in a mixed emotional state it should be verified that the facial (and possibly other) manifestation of it, even if it is not recognised as a blend of certain emotions, is perceived as a believable expression, one which may occur on real faces. Cunningham et al. (2003) have pointed out, by using video recordings, that such non-interpretable but believable expressions do occur on faces of real people.

2.2.4 Adaptation to the User Is it possible to tune the behaviour of the ECA according to (static or dynamic) characteristics of the user? Does the ECA maintain a model of the user, with aspects like expertise in the domain, age, gender, ethnicity, cultural and socio-economic background and personality? How is this acquired: by asking for the user profile, or by the ECA learning it? In what way does the user model influence the communication of the ECA (e.g., discourse strategy, being aware of safe or unsafe topics, what gestures should or should not be used in the users cultural context)?

Most of these aspects are far beyond the capacities of present-day ECAs, partly because of the lack of robust input possibilities (e.g., vision, voice analysis) in ECA applications to gain data about the user. The exploration of how single, static characteristics of the user influence her judgement of ECAs provides a basis for designing ECAs, to suit e.g., culturally different users the best. See the work of Isbister et al. (2000) discussed in section 2.4.3.

2.2.5 Discourse Capabilities An ECA may be more or less reactive. The extremes are the presenter and the pro-active conversational agent. In the first case, not only the content to be presented by the ECA, but all other information on the presentation is canned. In the latter case, the content of the presentation as well as meta-information on how to present it are generated on the fly, based on dynamically changing parameters of the conversation. These parameters may reflect aspects like emotional state, history of the conversation, status of task fulfilment.

Control How is the ECA controlled: by the application (in case of a presentation ECA), by the user (in case of most avatars in virtual forums), or by both (often the case for educational ECAs)? In the latter case, is there an explicit discourse model used; can the ECA display intention of turn-taking/turn-giving? How complex discourse patterns are allowed?

Is the ECA prepared to recover from erroneous input (content, timing), react to lack of input (after some time)? What modalities are used to indicate discourse states? Is feedback given to differentiate 'busy','idle', and 'waiting for input' states? Finally, how autonomous is the ECA, i.e. to what extent does it control itself? Does it take the initiative, for instance to signal a user that new information of interest has arrived?

According to Cassell and Thórisson (1999), non-verbal conversational signals of the ECA (e.g., averting gaze and lifting eyebrows when taking turn, performing beat gestures when providing content) are more valuable for the user than non-verbal emotional signals (e.g., smiling at the user). In their evaluation they used both objective measures of the users behaviour (e.g., number of hesitations) and subjective judgement by the user. Cassell and Vilhjálmsson (1999) have shown that in a chat environment avatars with autonomous non-verbal behaviour to express interest in conversing with others were considered more natural, more expressive and, interestingly, more easy to control, in contrast to avatars without any autonomous behaviour.

Input modalities of the user Though monitoring the reactions 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 should be perceived of the user.

2.3 Implementation Aspects

In order to be able to re-use and adapt an ECA, the technical requirements must be clear. Stating the technical parameters also helps to judge the design of the ECA independent of the limitations of the implementation or technical resources available.

For the ECA *body*, it is informative to know the modelling principle (polygon mesh or smooth surfaces, are textures used) and complexity of the model of the ECA (size of mesh). By what means was the model produced? Are there different levels of detail variants available? As of *non-verbal capabilities*, the quality (frame per second) of the rendered animation is relevant. The animation may have been designed by professional animators, or based on captured motion. As of *speech* and *natural language generation*, the external modules used (TTS, dialogue manager, NL generator) are relevant.

For judging the *conversational behaviour* of the ECA, the following implementation-related questions are important: In what form and detail should the relevant information be given? Is it in some standard format, like XML compliant markup tags? How long does it take to specify a typical input; what is the level of the input instructions which control the ECA's behaviour? How long does it take for the ECA to process these instructions, that is, to produce the final behaviour? It may be relevant to distinguish time spent on separate tasks (e.g., discourse management, generation of textual output, generation of speech).

The *operational requirements* may limit the applicability of the ECA. What software and hardware are necessary for using the ECA? What are the upgrading possibilities, considering hardware and software components used? What are the assumed operational parameters (e.g., noise in the environment, size of screen, data transfer mode, real-time versus off-line generation of output)?

2.4 Range of Applicability

The application context determines, by and large, what characteristics and ECA should have. As of application context, we distinguish presentation ECAs, information ECAs, educational ECAs, sales ECAs, entertainment ECAs and ECAs as research tool used to learn about (multimodal) communication. An ECA may be suitable as an interface for several examples of an application type (e.g., a talking head may read news items, weather reports, mails), or may be designed as a 'one-case' ECA specific for an application. Adaptability to different user groups depends on whether the ECA was designed in a parameterized and modular way. For instance, by providing access to the natural language, the non-verbal repertoire, the look, an ECA could be tailored for users of different cultures.

From a technical point of view, conformation to standards and modular design are relevant. Does the ECA body and animation conform to some standard (MPEG4, VRML)? Could it be re-used, can some of its aspects (e.g., look, accessories, body geometry) be modified? Can it be replaced by another model? Is it technically easy to modify or extend the repertoire of the ECA for each modality? Can the ECA be up/downgraded in terms of modality usage, e.g., according to the computer system capacity?

3. On Evaluation Methodology

Human-human communication and hence human-ECA communication is extremely complex, many parameters are involved, several of which are not clearly understood or, maybe, not even known. For instance, when finding a person nice, we (unconsciously) base our judgement on many aspects, such as look, way of speaking, gesturing, moving, usage of language. Hence evaluation work with different objectives is needed: to find out about the qualities of an ECA, compared to those common among humans, and to find out if an ECA has added value in a certain application context, and what is the best ECA for such a case. For the first case, the hidden assumption that ECAs should resemble humans, must be verified itself. In case of different applications, different aspects of the ECA may be relevant, and different users may have different expectations from and reactions to an ECA. In Chapter 9 by Catrambone et al. in this book, the importance of the nature of the application task is discussed in detail and illustrated by an experiment. When judging the merits of ECAs, the main issue is the identification of evaluation criteria, their interpretation and measurement. As these criteria involve responses (often subjective judgements) from the user, the criteria, and design rules abstracted from the evaluations, are more or less restricted in their scope of applicability. Finally, the collection and interpretation of empirical data should be done in a methodologically sound way.

In this section we address these issues briefly. First, we discuss the possible goals for ECA evaluation research, the types of design rules one may want to gain from the evaluations, and the relation of research on human-human communication to design and evaluation of ECAs. Then we address the problem of identification of evaluation criteria, in general. A sub-section is devoted to all the aspects of users which may influence their judgement of an ECA. Finally, we briefly sum up the sources and methods of collecting and evaluating data. For more in-depth discussion of doing evaluation research, see the Chapter 3 by Christoph in this volume.

3.1 Why to Evaluate?

A conscious setting of the goal is essential for the proper design of the evaluation and interpretation of the results. Basically, the target of evaluation is one of the following:

1 Find out the effect of single or multiple basic design parameters of the ECA on the perception and performance of the user (evaluation of the ECA itself). Specifically, the goal can be:

   (a) testing if a specific ECA fulfils some expectations;

   (b) finding out how to set certain parameters of the ECA to achieve some desired characteristics.

2 Find out about the merit of using ECAs for a given application (ECA as user interface evaluation). In this case too, one may be interested in:

   (a) testing if a specific ECA has added value;

   (b) investigating what ECA is the best for a given application.

Note that while the context differentiates the two cases, the sub-cases are similar in the sense that in case a) a concrete design has to be tested/verifified, while case b) requires exploration of the design field.

The first case corresponds to the micro-evaluation of ECAs, investigating the effect of certain modalities, the criteria to achieve a single characteristic (like intelligibility of speech, ability to indicate certain emotions). Testing if an ECA meets expectations (Case 1.a) is in particular relevant when ECAs are designed using artistic skills (at this date a common practice), and not explicit design guidelines. The required effect needs to be verified, as well as some additional, undesirable effects need to be excluded.

The second case corresponds to classical usability studies in human-computer interaction.

3.1.1 ECAs Like Humans? We tend to take it for granted that a good ECA should communicate as humans do. But this, as a basic design principle, needs to be verified itself. Namely, are we sure that humans will be 'fooled' to perceive a piece of moving object on the screen as a human being, with emotions and personality? There has been quite some evaluation work suggesting that the answer is yes. Surprising deceptions, associated with slight difference in the (static) facial expression (Sproull et al. (1996)) and human-like embodiment (King et al. (1996)) were reported. The extensive work by Nass and his colleagues led them to coin the 'computers as social actors' (CASA) hypothesis (see Reeves and Nass (1996)). It was shown that humans do perceive subtle differences in virtual characters, as voice characteristics, look, use of eyebrows, and interpret them similarly as they do in human-human communication. Bailenson et al. (2001) showed that people treat their own virtual alter-egos specially, in terms of reducing the size of the personal space respected around them. With other virtual characters, the distance patterns known from human-human communication were observed. Another sign of treating ECAs as humans is, when the user communicates with the ECA in an erroneous and somewhat messy way which is common in daily conversation. For instance, Cassell and Thórisson (1999) suggest that overlap between the user and the ECA talking can be interpreted not only as a dialogue error, but as a positive sign of the user taking the ECA as a real person, expecting him to interpret overlap in speech a sign of turn-taking intention.

These findings verify that we are on a good trajectory when making efforts to endow synthetic characters with embodiment and communicational traits used in human-human communication.

Ideally, the manifestation of the ECAs social role (e.g., salesperson, tutor, medical advisor) in behavioural and presentation parameters should be analysed with reference to sociological and anthropological investigation. The ECAs dialogues and behaviours should be compared to the role model. Isbister et al. (1998) give an example of such an evaluation of an agent playing the role of a bartender in a chat application. Prendinger and Ishizuka (2002) report on the social role effect (power relationship to user) of ECAs.

Based on the above finding, ECAs can also be used as research instruments for psycholinguists, psychologists and sociologists, to learn about the norms present in human-human communication.

But, of course, the above arguments do not imply that "the closer to realism the better an ECA is", neither that anthropomorphic ECAs should be the only possibility. Virtual characters have the potential of using additional, non-realistic cues, as has been demonstrated by the success of traditional animation characters. Moreover, the emphasis on non-realism possibly adjusts the expectations and frame of judgement of the user to a level more appropriate to the mental capabilities of the ECA.

3.1.2 Separation of the Application and the ECA The purpose of using an ECA is to provide better, or even a novel computer application. When interested in the added value of an ECA for an application, the base-line for verification, ideally, is the ECA-less version of the application. This, however, is much more problematic than with a traditional interface. The user has a single perception of a piece of software with an ECA embedded, and she might attribute aspects of the application (e.g., relevance of information provided, competence) to the agent's mental capabilities. For evaluation purposes it is important to separate what is the responsibility of the ECA and the underlying application, respectively. For instance, an ECA reacting with delays will be judged by users as unattractive and inefficient. But the cause of the delay can be very different:

1 It takes too long to generate the verbal and/or non-verbal signals to communicate the answer, promptly provided by the application.

2 It takes a long time for the application to produce the content of the answer (e.g., by searching a huge data resource), and the ECA is not prepared to inform the user that his answer is being produced by the application.

Clearly, in the first case the ECA is to blame. In the second case the essential cause of the delay is in the application. All the same, the ECA is still guilty, by missing a feature which could compensate for the inherent delay characteristic of the application. The deficiency may be on the mental level of the ECA: it might have a big expression repertoire, with the capability to indicate a processing state, but its view of the flow of communication is poor, not considering the processing state as one of interest for the user. But it can also happen that the ECA signals the processing state in a way which got misinterpreted or unnoticed by the user.

Moreover, the ECA technology allows entirely novel types of application which have not had an ECA-less counterpart, that is a system with traditional UI, because of the essence of the system is in the communicational capabilities of an ECA. It is impossible to imagine a version of the Erin the bartender system[1] with identical functionalities, but without an embodied bartender. In such cases it is interesting to compare the experience with the ECA with that of a real human in a similar role.

Another subtle point is that most of the current ECAs are designed for output. That is, the user is forced to communicate with the ECA by text input, requiring more time to perform and allowing for erroneous and irrelevant input. By introducing an ECA, on the output side the user interface is improved (at least that is the intention), on the input side it may become more cumbersome and error-prone.

3.1.3 Towards Design Guidelines An ultimate goal of evaluating ECAs is to produce design guidelines. Design guidelines may be of three kinds, depending on the cast of role of the independent/dependent variables.

- *ECA embodiment – performance* guidelines map embodiment parameters onto evaluation parameters (e.g., an ECA as a bank-clerk with formal dress is liked more than in a casual dress).
- *ECA mental aspect – performance* guidelines tell about a mental aspect parameter to be preferred for a certain performance objective (e.g., an ECA with extrovert speech will be liked more by extrovert users, than by introvert ones, see experiment by Nass and Lee (2000)).
- *ECA embodiment – mental aspects* guidelines tell how to choose some embodiment parameters to reflect the desired mental characteristic (e.g., Cowell and Stanney (2003) provide guidelines on how to achieve impression of trust by setting facial and gesturing characteristics).

Some design guidelines may be independent of the task and application context. E.g., to test the intelligibility of speech, depending on fine-tuning of speech synthesis parameters, and the effect of intonation, may be of general use for every application context. However, even such rules may need to be fine-tuned, with respect to application context: the ideal children story telling speech is surely different of the speech expected from a financial news reader.

For the applicability of rules mapping mental aspects to performance, or embodiment parameters to mental aspects, the application context is likely to be decisive.

Besides the above guidelines which are meant to specify aspects of an ECA directly, in order to meet some performance criteria, there are design guidelines which act on a higher level and express more complex relationships than mapping design and performance variables. Examples for such rules are: “The ECA should be consistent, that is, all relevant body and mental aspects should correspond to identical personality, social role, gender etc.”. “The ECA's personality should match the personality of the user.”

In case of all types of design rules, one should not forget about stating their scope, with respect to application type and user group. The rule on consistency is of general scope, valid for all applications and user groups. Contrary, the rule above telling that an ECA should have a personality matching that of the user, is of limited scope, and is applicable only to users who, in their human-human interaction are attracted by identical personalities. Isbister and Nass (to appear) conducted several experiments on this issue, and discuss the importance of all details in interpreting the results. One of the causes of the seemingly contradictory conclusions spread about ECAs is that conclusions and design guidelines are quoted without the scope of experimentation and applicability.

3.1.4 Evaluation for Evolution In an ideal software development scenario, evaluations are planned at different stages, to verify that the developed ECA fulfils expectations, or as a preliminary study to find out how users react to an ECA in the given application domain (see the Chapter 3 by Christoph in this book for more on evaluation at different stages). As the application of ECAs is still in its infancy, most of the evaluations are done on a small scale, at the place where the research has been carried out, to verify the potentials of the ECA technology. In the evaluations, especially when unexpected negative effects are experienced, it remains open to speculation if the effect is due to the design, the deficiency of some ECA components (like the quality of synthesized speech), to the incomparable measurements in this and other evaluations or to some hidden flaws in the methodology.

The design and implementation of an ECA should be an iterative process, where the next version is improved based on evaluations of the previous version, or of alternative versions. Some authors use the design-and-test loop concept, mainly to gain preliminary ideas about embodiment or test if some basic modalities function as assumed (e.g., if an implemented smile is recognized as such). We are not aware of long-term evaluations, except in a few cases of repeated experiments, to eliminate the novelty effect. The question arises if the results from one-session experiments carry over to situations where ECAs are used over months or years on a daily basis. It would be useful to hear about experience with mass applications developed by commercial companies, like Cantoche[2], Charamel[3], Headpedal (Griffin et al. (2003)) or sysis (see the Chapter 12 by Krenn et al. in this book).

In the future, the availability of design guidelines, with a clear scope of applicability, would make some (but not all!) evaluation stages superfluous.

3.2 How to Define the Evaluation Variables?

In an evaluation context, dimensions for judgement are to be selected, with corresponding variables of discrete or continuous values and methods to obtain these values from empirical data. In case of evaluating ECAs, one encounters major problems at all of the three stages: identifying the evaluation aspects, defining them in terms of measurable variables, and providing methods to measure them.

The origin of the problems is in the complexity of human-human communication. For instance, we often state that we like or trust somebody, but it is hard to find commonly accepted definitions of these natural-language terms (see e.g., the web site [4]). As discussed in section 3.1.3 with respect to the scope of design rules, trust may be different if it is to be applied to a bank clerk, or a game player. And there may be cases, like entertainment, when trust is of no relevance at all. Moreover, many synonyms and similar concepts are in use. In the ECA evaluation literature too, one encounters different working definitions of the same concepts, or similar definitions but given for different evaluation concepts. Sometimes there is no explicit definition at all, the concept is defined implicitly by the way it is measured. Several measurement techniques, like the most often used questionnaire, are based on lists and alternatives of further, fuzzy concepts, often made up for each study by the researcher. So one is puzzled how, for example, 'fun' in one study relates to 'likeability' in another? Or which possible sense of believability is meant: believable as a living entity, or as a believable action?

Furthermore, some evaluation criteria will clearly depend on certain perceived qualities of the ECA. E.g., likeability may depend on the personality of the ECA, on the intelligibility of its speech, etc. Some evaluation criteria may not be completely independent of each other. For example, showing friendship may be an important aspect of inducing trust in the user and thus enhance the usability of an electronic commerce application. In our discussion we will refer to *high-level*, or *compound evaluation criteria* which involve others, as opposed to *low-level*, or *basic evaluation criteria*.

So when setting the evaluation aspects, the following two choices are to be made explicit:

1 Which (objective or subjective) evaluation criteria are of interest?

2 How are these interpreted, related to each other and to the qualities of the ECA?

Ideally, it is the task of experts in psychology, sociology and of the application domain to identify what aspects are relevant in certain application scenarios. Sessions with ECAs can serve as experimental settings to find out also about these aspects. What the main task for the ECA evaluator is, is to find out how to decompose a subjective aspect (e.g., likeability) into aspects which can be related to ECA design. Such aspects may be some objective performance aspects (e.g., understanding well what the ECA says), may concern mental aspects of the ECA (personality judgement) or some aspects of embodiment (e.g., gender or aesthetic appeal). These composite factors (not only their values!) may differ in different contexts. Hence one should be careful when defining a concept a priori in terms of others, without giving verification rooted in the application context.

This observation leads to the methodological deficiency in measuring these fuzzy concepts. Mostly, a set of questions are bunched together as measurements for one concept, without any verification of using just those and not other questions, and the way of gaining a single measurement value (e.g., by averaging) based on the answers to the questions. In particular, the relation between the definition of the concept and the way it is measured remains unclear. It remains problematic if the data obtained by the measurement are valid, in particular when a psychological construct is evaluated (see the Chapter 3 by Christoph in this book for the discussion of validity). One should make clear the relationship between collected data and evaluation variables. Two different approaches are used.

One possibility (and common practice), as discussed in the Chapter 3 by Christoph in this book in detail, is to define a complex evaluation variable in terms of its (directly measurable) phenomena in advance. For basic variables, the measurement method may be widely accepted; for instance arousal can be measured by blood pressure, or by observing facial expressions. For compound variables there may be measurement methods used in psychology, like for instance the desert survival problem to judge trust, or different methods to test intelligence. Note that in this case an alternative of measurement methods is offered, and it depends on the situation which one(s) to use.

For compound variables which do not yet have accepted measurement methods, in the previous sense, another, exploratory approach can be used. The compound variable is decomposed into simpler, independent components, which each get measured. For instance, for liking, arousal may be one component, subjective judgement of appeal another, helpfulness (in itself a complex variable) yet another. From these measurements, an aggregate value for the high-level variable is derived, e.g., by averaging (as is mostly done), or by some more subtle partial comparison of the measurement results for the components. Note that in this case all components need to be measured. It is possible that for a component well-proven alternative measurement methods are available, as explained earlier.

The identification of the components of a high-level evaluation variable is a non-trivial task in itself, as explained earlier. At the present state of the art of ECA technology and evaluation, an unbiased, mathematical approach, as used by, for instance, Nass et al. (2000) and Cowell et al. (2003), seems to be the most appropriate for us to learn also about the relevant evaluation aspects and their relationship.

It would be an interesting research topic to identify some categories of tasks or application contexts, and provide some objective definition and measurement methods for the relevant evaluation criteria, by using mathematical methods for decomposing the concepts into components, and established methods to measure those. This would produce a common ground for evaluating ECAs to be used in the same application context.

For the measurement of evaluation variables, expertise in related fields could be used. There are examples of adopting psychological tests and case-problems to judge perceived personality and trust (even in an application independent sense!). One could consult experts to forge a new measurement method. Moundridou and Virvou (2002) have asked 15 classroom tutors to come up with measurement for attention of students communicating with a tutoring ECA.

3.3 Testing by what Users?

When performing usability tests, the group of subjects should be selected carefully, as users with different characteristics may interact with and judge an ECA differently. Below we outline the aspects which may be relevant for ECA users, and refer to findings obtained so far.

3.3.1 Demographic Data The following demographic aspects of the user may be relevant for the ECA usage: gender, age, fluency in the language of communication, ethnicity, computer skills and familiarity with ECA technology.

Gender Most of the evaluation studies are aware of the potential importance of gender, as the gender distribution of the subjects is almost always reported. However, the results are still scattered and sometimes contradictory, so not sufficient to formulate design guidelines with respect to the gender of the users. Comparing different designs of full embodied agents for a retail application, McBreen et al. (2000) suggest that females may prefer to interact with agents of their own gender. Buisine at al. (2003) did not find such a correlation between the gender of the ECA and of the user, but reported on gender difference in preference for different non-verbal strategies of the ECA.

Age The age of the user has hardly been considered as an influential factor for ECA evaluation. Most often it is assumed that the experimental subject's age (usually student age) is the same as the target groups age. It would be interesting to investigate how age influences preferences for the looks and communicational modalities of an ECA. Describing an emotional expression model for chatterbots, Paradiso and L'Abbate (2001) stated that it was important to take into consideration the age of the user as the expressiveness of an agent should be stronger for younger users.

Ethnicity Ethnicity is meant to indicate the ethnicity of a person as visible from her looks. Studying the effect of the ECA ethnicity, Nass et al. (2000) found that when ethnicity of users and agents matched, the ECA was regarded as socially more attractive.

Language In the evaluation literature authors describe the level of language knowledge of the experimental subjects in a variety of terms, like "first language is English", "fluent English", or implied by being "3rd year student at American university X". The precise characterisation of the level of the communicational language skills may be relevant when mental aspects of the agent are to be judged, or if efficiency is measured by recall or task performance.

Computer skills One may expect that users who know more about the mechanisms of computer applications and have a high proficiency in using computers perceive ECAs as less attractive, and also gain less (or even lose) in efficiency by using them. All the same, small-scale evaluations are often done by computer science students, who cannot be considered as good representatives of an intended user group. Proficiency in using computers should be established on the basis of a series of factual questions concerning using computer at work and in private life, as done e.g., by Cowell et al. (2003).

Familiarity with ECA technology It is often mentioned that the 'novelty effect' biases a user's judgement. In practice, almost all test experiments are prone to this effect. On the other hand, the testing subjects should not know more about the ECA technology than the intended user group, as people from the ECA research field might be biased in their judgement and skilled in using ECAs.

3.3.2 Psychological Data The mental characteristics of the user (other than language skills) are surely reflected in his preferences for ECAs. Research on what the relevant user characteristics are, and how they should be taken into account when designing an ECA, is still in its infancy.

Personality The following personality characteristics of the user have been considered in the context of ECA evaluation: self-esteem, introversion/extroversion, and locus of control. Resnick and Lammers (2000) showed that users with low or high self-esteem reacted differently to error messages. Studying trust through relational conversational strategies, Cassell and Bickmore (2001) claimed that social dialogue had a positive effect on trust for users with a disposition to be extroverts. Nass et al. (2000) found that individuals had more fun with agents whose non-verbal cues matched their own personality. Rickenberg and Reeves (2000) showed that the locus of control of the user was relevant in the anxiety evoked by the ECA (see section 4.2.5 for discussion).

Affect intensity Affect intensity is used in psychology to characterise the intensity of emotional response of the user to a given level of affect stimulus (Larsen (1987)). It appears that high affect intensity individuals, when exposed to emotional stimuli, produce more affect related cognitive responses as well as experience stronger emotional reactions. Thus users emotional reactions to an ECA as well as their preferences for certain types of ECAs could be related to their emotional profile or level of affect intensity.

Cognitive style Cognitive style is the collection of stable aspects of how people organize their thoughts, deal with sensory input and communicate ideas. In HCI cognitive style has been used as a common entry of user profiles (Benbasat et al. (1981)). Modelling user cognitive style might be particularly relevant for pedagogical agents. User cognitive style might also influence their preferences for specific styles of multi-modal communication in other applications.

Perception and body capabilities Among the users there may be ones who have deficiencies in using some communicational channels. Hearing and the capability to read faces have been used to pre-test users to exclude anomalies. Handedness of the user may have consequences on the judgement of gesturing of an ECA, especially in case of instructional tutoring applications.

3.3.3 Culture A culture's impact on a person is to be noticed in his communication, norms and beliefs (de Rosis et al. (2001)) and behaviour (Hofstede (1997)). Thus culture should also be an aspect of the user profile, when designing an ECA to be used by a multicultural public, e.g., via the web. Isbister et al. (2000) examined the effect of the agent on crosscultural communication. They found that two cultural groups with very different interaction styles and norms; namely American and Japanese had different impressions on the same agent and they reacted in different ways.

While ethnicity can be decided at a glance, there is much discrepancy in how to elicit the user's culture (and what is meant by it). One comes across cultural descriptions like "CS student of Chinese origin" or "with Western/Eastern philosophical tradition" (King et al. (1996)). It requires further research to provide methodology to set useful categories of culture. Will a student, fluent in English and having spent 10 years in the USA, perceive and judge the non-verbal gestures of an ECA in a similar way as an American born student?

3.4 How to Collect and Evaluate Data?

Once it is clear what aspects of an ECA are to be evaluated and in what context, one has to design a setting for collecting relevant data, and a way of interpreting them. Chapter 3 by Christoph in this book is devoted entirely to evaluation methodology, in this section we give a brief summary of the most important issues.

3.4.1 Empirical Data Collection According to Dehn and Van Mulken (2000), an ECA may have influence at three levels:

- the user's behaviour during interaction,
- the user's subjective perception of the interaction, and
- longer term effects on the user.

The effects at the three levels do not always coincide, the subjective perception of the user may differ from the conclusions based on observation or testing the final outcome. Höök et al. (2000) evaluated their Agneta and Frida system, and noticed that the same user who was often smiling while interacting with the system, did not like the characters, according to the post-session questionnaire on subjective impressions.

Below we list the data collection methods most appropriate for evaluating ECAs.

Observation of user behaviour takes place at the work-place or in the laboratory, in order to get basic impressions of ECA usage.

Experiment is used for a systematic evaluation of ECA designs or elicitation of characteristics of human-human communication.

Benchmarks and comparative tests are standardized forms of experimental procedures, based on carefully constructed standard tasks. It is still a challenge to define benchmark scenarios to test different aspects of ECAs in an application independent way, as a function of certain characteristics of the ECA. Choice Dilemma Situations and the Desert Survival Problem have been used by Nass et al. (2000) to test the effect of ECAs in an way independent of the domain of the application. Recall rate can be used for testing the learning effect in arbitrary domains.

Questionnaire and interview are done with paper-and-pencil, and face to face, respectively. It is known that the interview technique may bias the subjects answers.

Usage data provides some quantitative characteristics of interaction of the user, based on logged users input action or recorded non-verbal behaviour like eye gaze or head movements (registered automatically during the entire session with the system).

Biomedical data are gained by measuring directly some biomedical characteristics of the user during the interaction. Blood pressure and skin conductivity have been used by Mori et al. (2003) to get an objective picture of affect arousal of the user during the entire interaction.

3.4.2 Interpreting the Data The interpretation of collected data may be a source of flaws for the experiment. Mistakes range from misusing the data (e.g., misjudging them as indicators for some phenomena) to the incorrect use of statistical methods. In a nutshell, the following major points must be taken care of:

- Most often, it is decided a priori that one measurement dimension, or the average of several different measurement dimensions is used as the value for an evaluation variable. Such an approach implicitly determines the evaluation aspect in terms of the measurement data. Verification of such an implied definition should be given, by referring to common practice or to some theoretical foundation. If these are not available, the motivation for the chosen measurements and mapping should be stated.

- When it is not yet well established what the evaluation dimensions should be, exploratory data analysis should be considered. Instead of an a priori interpretation framework, correlation between different data (e.g., answers to different questions) should be investigated by some sound method, like principle component analysis. By such an approach one can derive a few characteristic compound dimensions (consisting of sub-variables corresponding the specific aspects, each represented by a separate entry in the initial data collection) to judge the ECA.

- Simple comparison of numerical data or respective averages is sufficient only as descriptive evaluation.

- To draw conclusions, statistical tests are to be conducted, with carefully chosen and documented parameters.

- To perform specific statistical tests, data from a sufficient number of properly selected users are needed.

- Certain user characteristics might have a discriminative effect, which should be checked.

- If user observation data are labelled by evaluators, care should be taken that the labelling is correct (usually by using multiple evaluators and ensuring agreement between their judgments).

4. Dimensions of Evaluation

In this section we identify aspects that, in our view, are most relevant for evaluation. We provide a definition for each evaluation criterion. Then we discuss the different usage and measurement of the concept in the literature.

In the first subsection, we deal with the aspects which are strictly related to the performance of an ECA as an interface. The performance-related aspects can be evaluated basically by objective measures of behaviour and results achieved. As 'good performance' is beneficial in all application domains and for all users, these objectives are universal, though the 'good performance' may have different meaning for different applications.

In the second subsection, we turn to the issue of the users experience with an ECA. The corresponding evaluation criteria are subjective and are more difficult to measure. Furthermore, it depends on the application domain and the user group, which of the possible qualities perceived are relevant for the ECA.

4.1 Usability

As a starting point for usability, we refer to the concept as described by Nielsen (1993) for general HCI. In this section on usability we discuss the *task performance* dimensions, namely *learnability*, *efficiency*, *memorizability* and *error*. Nielsens fifth category, satisfaction goes to the next section in a strongly modified form, as a dimension of user perception. In this way we separate the evaluation criteria related to objective performance and to subjective perception of the user.

In HCI, there are generally accepted heuristic guidelines to judge an interface. For instance, in order to judge the consistency of a user interface, the use of shortcuts, menus and other selection and navigation devices, layout and colours should be looked at and compared with common practice in other applications as well as multiple use in the given application. One can spot easily if, for instance, the usage of red colour or the shortcut key Ctrl-C are not consistent with common practice. So a 'quick and dirty' evaluation of a traditional user interface can be done by checking heuristic design rules.

In case of ECAs, we do not have yet such a complete and fixed set of heuristic design rules. The suggestion by Sanders and Scholtz (2000) provides rules to judge the natural language dialogue capabilities of an ECA. Many of the objectives stated in the heuristics for traditional user interface design are very likely desirable also when the interface is an ECA, though this has to be verified. The major problem is to be able to tell if a given ECA fulfils a requirement. Using the previous example, the question of consistency of an ECA is a far more complex issue than that of a traditional user interface. It involves the subtle correspondence of almost all design aspects. As discussed before, the identification of the evaluation criteria as well as the realization of the desired effects, in terms of the design parameters of the ECA, are open issues themselves, asking for multidisciplinary research.

This is the reason why in ECA evaluation the method of heuristic evaluation conducted by experts is hardly present, but empirical tests (also used in testing traditional software) are more often performed. In different, designed scenarios test subjects interact with the ECA. In order to measure the usability concepts as dependent parameters, data sets are collected and metrics are developed.

4.1.1 Learnability, Memorizability and Ease of Use

Learnability is the easiness/difficulty of figuring out how the ECA 'works', from the point of view of maintaining a discourse with it. *Memorizability* is to express how easy it is for users to remember those interaction strategies. *Ease of use* is a compound criterion, consisting of learnability and memorizability

The main motivation of having an ECA as an interface is its ease of use. Ideally, the user communicates with an ECA just as she does with a real person. In this ideal case, there is hardly anything to be learnt, as the user has been practising the type of natural communication in his daily life. As in practice ECAs are far from full-fledged humans in their communicational means, there are several concerns to judge learnability: Are users provided with sufficient instructions to understand how to interact with the ECA? Does the ECA tell, by way of introduction or when appropriate, what his limitations and powers are? Are the agent's limitations and capabilities (communication and mind) clear from his behaviour, or are wrong expectations generated? How natural it is (compared with human-human interaction) to communicate with the ECA?

Memorizability is quite important for novice users. Actually memorizability plays a role as a factor in learnability too. If some steps in the interaction process are hard to memorize, this, of course, hinders learning.

4.1.2 Efficiency *Efficiency* is the relation between the success (accuracy and completeness) in achieving certain goals and the mental resources and time spent on it.

Efficiency can be defined as the degree to which the ECA enables the task to be completed in an effective and economical fashion. Depending on the kind of task, efficiency has different measures. When there is a clear-cut task which gets either performed or not (e.g., booking a flight), the number of goals/tasks achieved in a period of time, or the time needed to complete the task, can be measured. In order not to consider the extra time spent with the ECA 'for its sake', the *on-task time*, devoted to solving the task, should be considered. For other domains (e.g., learning), task fulfilment quality must be evaluated (e.g., by comprehension or recall). As of *mental resources*, low-level components like fatigue, stress and perceived mental load are measured. Stress and mental load relate the concept to perceived task difficulty (discussed in 4.2.6).

Apart from evaluating task performance efficiency in an ECA application one could evaluate the efficiency of the ECA's communicative functions by itself. For instance, the ECA's communicative skills are a general property which could be evaluated by experts separately from a specific application context. One could also design experiments to evaluate them, possibly using a context tuned to this evaluation purpose.

4.1.3 Errors *Errors* indicate the relative amount and type of mistakes occurring while interacting with the ECA.

Common error categories can be identified, such as: misunderstanding (as of information content) of the ECA by the user or of the user by the ECA; problems in the dialogue management (whose turn it is, is the ECA idling or still active, deadlock situation). The relative number of occurrences of different types of error, as well as relative time spent on recovering from them, are good indicators how error-prone the ECA is. A related issue is whether the ECA provides active help for the user to recover from errors by, for instance, asking to repeat her input, or taking the initiative to recover from interaction errors.

4.2 Evaluation of User Perception of ECAs

In this section we discuss evaluation aspects of the ECA which essentially have to do with the perception of the user. Some of the aspects have a corresponding or related usability dimension (like satisfaction and usability), others like engagement and trust can be measured by observing the behaviour as well as by questioning the user.

4.2.1 Satisfaction *User satisfaction* is the perception by the user that her interaction with the ECA serves ones intentions in a rewarding and agreeable way.

Though one of the most measured aspects, user satisfaction is a difficult concept. It is related to objective usability: if an ECA is inefficient and difficult to use, the user will be, basically, unhappy with using it. However, it has been suggested by several experiments (see discussion in 4.2.6), that the subjective impression may deviate from the objective performance: an ECA makes the user perceive the interaction and even the quality of the service more positive. This so-called ‘persona effect’ is another major motivation of applying ECAs.

A user’s reactions to an ECA depend on several subjective factors, such as the importance of achieving some goal with the application, her (positive or negative) prejudice of the outcome of using the ECAs. A weakness of many of the experiments with test subjects is that the situation is not ‘real’, there are no consequences like passing or not an exam after sessions with a tutoring ECA, or gaining or losing money when following the advice of a broker ECA.

User satisfaction is a vague and in itself multi-dimensional concept, with possible components like *emotional liking* and *arousal*, assessment of *attractiveness*. It depends very much on the user’s own characteristics, what is ‘attractive’ and ‘pleasant’ for her. In order to get insight into the factors of user satisfaction, one should carry on evaluation research where the possible dimensions of the concept and the user profile are taken into account, as discussed in Section 3.

For instance, Nass et al. (2000), investigating the consequences of ECA’s personality, use fun as one of the concepts that indicates user satisfaction. Based on a factor analysis of responses to a questionnaire they define fun (triggered by using an ECA) as a high-level concept of the following components: enjoyable, exciting, funny and satisfying. Thus, instead of just evaluating for ‘pleasant versus unpleasant’, a more complete model of emotion is needed to cover all aspects of emotional assessment. Such a model is the pleasure, arousal dominance model described by Mehrabian and Russell (1977). Evaluating user experiences should include evaluating all relevant emotional as well as social aspects of the ECA.

4.2.2 Engagement An *engaging (involving, appealing)* ECA motivates the user to spend time dealing with it while perceiving the activity as pleasurable in itself. Engagement is a high-level experience dimension. Its constituents depend on the application context. For instance, for an ECA as personal assistant these include likeability and trust, but if the ECA acts as tourist guide, these factors are less decisive, but the level of entertainment is. In our sense, engagement is even stronger than user satisfaction. One cannot be engaged by an ECA if one does not feel satisfaction while interacting with him. On the other hand, one can be satisfied with a non-engaging ECA, for instance when the ECA helps with a task one does not like doing, but has to do. In that case one will not be inclined, in spite of the satisfaction, to spend more time with the ECA. In the literature, this point of view is not always shared. Koda and Maes (1996) treat satisfaction on the same level as likeability and intelligence.

Both the relevance of the services of the ECA and its design aspects (look of the body, gesturing and speech, personality) have an effect on engagement. A correlation between user personality and ECA engagement is reported by Bickmore and Cassell (2000). Active users (who take the initiative in talking to the ECA) found the estate-agent REA more engaging when she got them involved in small talk, passive users when there was no small talk.

4.2.3 Helpfulness An ECA is *helpful*, if in the users perception the ECA behaves in a cooperative way to assist her in achieving her goals and in resolving difficulties.

A way to paraphrase the definition is that the ECA should behave as a good assistant. Obviously, the perception of helpfulness is related to a large extent to usability aspects like how, when and what information the EAC presents. But less obviously, the ECA's visual design characteristics also play a role. For instance, Lester et al. (1997) conducted experiments with an educational ECA giving advice on two different levels (principle-based or task specific), with or without instructional animations. Subjects rated the ECA version with principle-based advice, demonstrated by animations as significantly the most helpful. McBreen et al. (2000a) evaluated retail agents, where the controlled variables were gender and visual characteristics. As for helpfulness they found that the fully realistic (video) head scored higher than the 3D talking heads. What was more interesting, that the *male* 3D talking heads (also with male voice) scored even lower than either of the stills. The last result is difficult to interpret for a reason which is paradigmatic for a general problem with this kind of evaluations. When manipulating gender in this example, more than one (high level) parameters are manipulated: visual characteristics and voice. This is unavoidable because the ECA has to be consistent!

4.2.4 Naturalness and Believability An ECA is *natural (life-like)*, if it is in line with the expectations of the user about a living, acting creature with respect to its embodiment and communicative behaviours. When on top of that its task performance is perceived as meeting the expectations, it is *believable (credible)*.

The user judges the ECA based on its look and communicational behaviour. These should be consistent at each moment and at different points of time. They all should give the impression of a real living creature. For instance, a robot-like voice, or the lack of idle motion destroy the illusion of life, and thus, naturalness. Furthermore, consistency with the domain the ECA functions in is expected too: information should be provided in such a way that the user is willing to take the information seriously. Believability in this sense is not equal to taking the ECA as *real*. ECAs often have deliberately a non-realistic design, with non-realistic features. Not only the (yet significant) shortcomings of the technology do not allow to produce perfect clones of real humans, the realm of non-realism has additional advantages, like the enhanced expressivity of cartoons. But in case of non-realistic embodiment too, believability is an important evaluation criterion.

We found two examples of evaluations of believability, where the concept was used in accordance with our definition. In a teaching application Lester et al. (1997) ask test subjects: “Did you believe the advice you got from Herman the bug (the teaching agent)?” which we interpret as: “Did you take the advice by Herman as an advice given by a teacher?” In the literature on ECA evaluation, the distinction between believability, trust and credibility is sometimes quite unclear. For instance Nash and Lee (2000) talk about *voice credibility* (for a synthetic voice of a reviewer) as a high-level concept composed of the following three qualities: credibility, reliability and trustworthiness. Another evaluation concept was credibility of the ECA (in the role of a book reviewer), which was measured by a standardized trust scale.

4.2.5 Trust *Trust (credibility)* is the belief that the ECA has benevolent intentions towards the user and has the competence to put those into effect.

Cassell and Bickmore (2000) further differentiate trust: “A useful distinction can be made between a cognitive state of trust and trusting behaviours. Trusting behaviours involve making oneself vulnerable to other people in any one of a number of ways.” In their experiment the same authors provided evidence (Cassell and Bickmore (2001)), that the users subjective statement about her trust in the ECA does not necessarily coincide with a trusting behaviour towards the ECA. They also showed that small talk increased the trust in the agent, but among extrovert users only. Cowell et al. (2003) have reported on the correlation of perceived trust and task performance, as well as other perceived qualities of the ECA.

Rickenberg and Reeves (2000) tested the reaction of subjects (distinguished on the internal/external locus of control dimension) who had to perform tasks on web sites in the presence of an ECA which behaved as if monitoring the user. Monitoring produced anxiety especially for subjects with external locus of control, yet in the monitoring condition subjects trusted the website more than the same website without an ECA.

In McBreen et al. (2001) users report not having trust in an e-banking application, featuring an ECA because “they have not enough confidence in the technology yet”. Given the accepted status of e-banking, this result points at the danger of adding an immature ECA on top of a proven application.

4.2.6 Perceived Task Difficulty *Perceived task difficulty* is the subjective judgement of the difficulty of the task.

This is one of the parameters referred to in the discussion about the persona effect, initially coined by Lester et al. (1997). Namely, that the presence of an ECA makes the user perceive the task as easier, without a measurable difference in task performance. Such effect has been reported with tutoring systems in different domains, like the operation of a pulley system by Van Mulken et al (1998), linear equations by Moundridou and Virvou (2002) or biology by Lester et al. (1997). Recently, Baylor (2003) conducted an experiment which suggests the superiority of the ‘split-persona effect’: having two separate pedagogical agents with different roles improved learning performance and perceived value of the agents. On the other hand, the experiments by Van Mulken et al. (1998) suggest that the assistance of an ECA has no effect on short-term learning; moreover there was no persona-effect at all in case of the less technical application which dealt with photos of human faces.

4.2.7 Likeability An ECA is perceived as *likeable* (sympathetic) if the user feels positive about (some of) its traits and behaviours.

Likeability is a compound concept too. In a loose sense, it is the judgement of the ECA, also its personality. As this judgement is user-dependent, one should not equal likeability to a kind, friendly personality of the ECA. (Think of how one can dislike a ‘keep smiling character too.) Moreover, additional design (like look) and perceived aspects (helpfulness, trustworthiness) of the ECA also play a role. A difference with respect to engagement is that the ECAs competence as a task performer and the relevance of the performed task do not enter here. In this case too we encounter terminology problems in the literature. Buisine et al. (2003) and Koda and Maes (1996) evaluate likeability by directly including the term in a questionnaire. Rickenberg and Reeves (2000) on the other hand used a compound concept (derived by factor analysis of a sixteen-item questionnaire), containing items like enjoyment, fun and boredom.

4.2.8 Entertainment An *entertaining* ECA is amusing in a non-task related way, thereby making performing the task more agreeable for the user.

Both the relevance of the services of the ECA and its design aspects (look of the body, gesturing and speech, personality) have an effect on engagement. Van Mulken et al. (1998) report on a technical explanation (of a pulley system) and a non-technical recall task (remembering data on new employees) presented with and without an ECA. In the technical case the explanation with the ECA was judged significantly more entertaining. No difference was found for the non-technical task. Although the authors are not sure how to interpret this difference, to the ECA designer it shows the importance of the application context.

5. Conclusions

We proposed a framework for comparing and evaluating ECAs. We introduced the general and most important issues one has to take care of when starting research on evaluating ECAs. We discussed the design parameters of ECAs in detail. Then we took a critical look at the relevant literature to elicit common terminology of evaluation aspects. While we did our best to provide a complete list and acceptable working definitions for fuzzy concepts which have been used widely and controversially, we do not claim that our list is closed. Just the opposite, we will be happy if our work will induce some debate and will lead to improvements and extensions on evaluation aspects. In our discussions we emphasized the proliferation in methodology. The next step is to settle some methods (till the detail of questions to be asked) and provide benchmarks as the standard way to evaluate certain aspects of ECAs.

Our secondary goal was to draw attention to the necessity of a common framework. In our view, a common reference framework will facilitate many tasks in the ECA community:

- To compare ECAs, from a design and technical point of view;
- to facilitate the re-use and adaptation of existing ECAs;
- to help researchers doing evaluation to converge to some design guidelines;
- to point out ‘white spots’ in human-human communication, and in ECA evaluation.

We claim that by taking a systematic and critical look at design categories, evaluation criteria and evaluation methods, the research efforts can be spent better. Not only on a short term, by avoiding pitfalls of making unsound conclusions or developing superfluous features. But also on a long term, by having a clearer view within the research community, and presenting a, maybe, more subtle but sound and not less challenging image of our field for the outside world about what has been achieved and what we are after.

But with the near future in mind we want to conclude with the following concrete recommendations to researchers in this field:

- Keep in mind that all the design parameters mentioned in section 2 (and possibly further ones) may influence the impact of your ECA in often yet unknown ways.

- When trying to find design guidelines, vary only one of the parameters at a time, i.e. comparing a 2D cartoon to a 3D cartoon and a 3D cartoon to a 3D realistic ECA is more instructive than comparing a 2D cartoon and a 3D realistic one. The latter kind of evaluation makes sense in practical cases only, where two alternatives to choose from are available, but it does not lead to general design guidelines.

- When evaluating an ECA with an application, take care to separate the effect of the two (see 3.2) if you want to draw conclusions on the effect of the ECA. When possible at all, use the application without an ECA as baseline.

- Ask yourself what the intended user group is. Take demographic data and user characteristics into consideration (see 3.4).

- Whenever possible, use evaluation dimensions and measurement methods also used by others. When not possible, discuss why and define them.

- Lets try to reach agreement on evaluation (especially user perception) dimensions, their definition and measurement method in order to leave behind us the incompatibility problems discussed in section 3.3

Notes

[1] www.extempo.com

[2] www.cantoche.com

[3] www.charamel.com

[4] http://dict.die.net/trust/ for several alternatives

References

André, E. and Rist, T. (2000). Presenting Through Performing: On the Use of Multiple Lifelike Characters in Knowledge-Based Presentation Systems. In *Proc. of the Second International Conference on Intelligent User Interfaces*, pp. 1–8.New Orleans, Louisiana.

Bailenson, J. N., Beall, A. C., Blascovich, J., Raimundo, M. and Weisbuch, M., (2001). Intelligent Agents Who Wear Your Face: Users Reactions to the Virtual Self. In De Antonio, A., Aylett, R., Ballin, D., editors, In *Proc. of the Third International Workshop Intelligent Virtual Agents, IVA 2001*, pp. 86–99, Madrid, Spain, Lecture Notes in Computer Science 2190, Springer.

Barker, T. (2003). The Illusion of Life Revisited. In *Proc. of the AAMAS03 Workshop on Embodied Conversational Characters as Individuals*, Melbourne, Australia.

Baylor, A. L. (2003). The Split-Persona Effect with Pedagogical Agents. In *Proc. of the AAMAS03 Workshop on Embodied Conversational Characters as Individuals*, Melbourne, Australia.

Benbasat, I., Dexter, A. and Masulis, P. (1981). An Experimental Study of the Human/Computer Interface. *Communication of the ACM*, 24(11): 752–762.

Bickmore, T. and, Cassell, J. (2001). A Relational Agent: A Model and Implementation of Building User Trust. In *Proceedings of the CHI'01.* pp. 396–403.Seattle, Washington.

Buisine, S., Abrilian, S., Rendu, C., and Martin, J.C. (2003). Evaluation of Individual Multimodal Behavior of 2D Embodied Agents in Presentation Tasks, In *Proc. of the AAMAS03 Workshop on Embodied Conversational Characters as Individuals*, Melbourne, Australia.

Cassell, J. and Bickmore, T. (2000). External Manifestations of Trustworthiness in the Interface. *Communications of the ACM*, 43(12): 50–56.

Cassell, J. 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.

Cassell, J. and Vilhjálmsson, H. (1999). Fully Embodied Conversational Avatars: Making Communicative Behaviors Autonomous. *Autonomous Agents and Multi-Agent Systems*, 2(1): 45–64.

Cowell, A.J. and Stanney, K.M. (2003). On Manipulating Nonverbal Interaction Style to Increase Anthropomorphic Computer Character Credibility. In *Proc. of the AAMAS03 Workshop on Embodied Conversational Characters as Individuals*, Melbourne, Australia.

Dehn, D. and Van Mulken, S. (2000). The impact of animated interface agents: a review of empirical research. *Int. Journal of Human-Computer Studies.* 52: 1–22.

De Rosis, F., Pelachaud, C. and Poggi, I. (to appear). Transcultural believability in embodied agents: A matter of consistent adaptation. In Payr, S. and Trappl, R., editors, *Agent Culture: Designing Human-Agent Interaction in a Multicultural World.* Laurence Erlbaum Associates, New York.

Dryer, D. (1999). Getting personal with computers. *Applied Artificial Intelligence.* 13: 273–295.

Griffin, P., Hodgson, P., and Prevost, S. (2003). Character User Interfaces for Commercial Applications. In *Proc. of the AAMAS03 Workshop on Embodied Conversational Characters as Individuals*, Melbourne, Australia.

Hofstede, G. (1997). *Cultures and Organisations: Software of the Mind.* McGraw-Hill, New York.

Höök, K., Persson, P., and Sjölinder, M. (2000). Evaluating Users Experience of a Character Enhanced Information Space, *Journal of AI Communications*, 13(3): 195 – 212.

Isbister, K. and Doyle, P. (2002). Design and Evaluation of Embodied Conversational Agents: A Proposed Taxonomy. In *Proc. of the AAMAS02 Workshop on Embodied Conversational Agents: Lets Specify and Compare Them!* , Bologna, Italy.

Isbister, K. and Hayes-Roth, B. (1998). *Social Implications of Using Synthetic Characters: An Examination of a Role-Specific Intelligent Agent.* KSL-98-01 Stanford, Knowledge Systems Laboratory.

Isbister, K., Nakanishi, H., Ishida, T., and Nass, C. (2000). Helper Agent: Designing an Assistant for Human-Human Interaction in a Virtual Meeting Space. In *Proceedings of the CHI 2000*, pp. 57–64. The Hague, The Netherlands.

Isbister, K. and Nass, C. (2000). Consistency of Personality in Interactive Characters: Verbal Cues, Non-Verbal Cues, and User Characteristics. *International Journal of Human-Computer Studies*, 53: 251–267.

Isla, D. and Blumberg, B. (2002). Object Persistence for Synthetic Creatures. In *Proc. of AAMAS02*, pp. 1356–1363.

King, J. and Ohya, J. (1996). The Representation of Agents: Anthropomorphism, Agency, and Intelligence. In *Proc. CHI96*, pp. 289–290.

Koda, T. and Maes, P. (1996). Agents With Faces: The Effects of Personification of Agents. In *Proc. of HCI'96*, pp. 98–103. London, UK.

Larsen, R.J. and Diener, E. (1987). Affect Intensity as an Individual Difference Characteristic: A review. *Journal of Research in Personality*, 21: 1–39.

Lester, J., Converse, S., Kahler, S., Barlow, S., Stone, B., and Bhogal, R. (1997). The Persona Effect: Affective Impact of Animated Pedagogical Agents. In *Proc. of CHI97*, pp. 359–366, ACM Press, New York.

Massaro, D. (1998). *Perceiving Talking Faces.* The MIT Press, Cambridge, MA, USA

McBreen, H.M., Shade, P., Jack, M.A., and Wyard, P.J. (2000). Experimental Assessment of the Effectiveness of Synthetic Personae for Multi-Modal E-Retail Applications. In *Proc. of Fourth International Conference on Autonomous Agents*, pp.39–45.

McBreen, H. M., Anderson, J., and Jack, M. (2001). Evaluating 3D Embodied Conversational Agents in Contrasting VRML Retail Applications. In *Proc. of AAMAS01 Workshop on Representing, Annotating, and Evaluating Non-Verbal and Verbal Communicative Acts to Achieve Contextual Embodied Agents*, Montreal, Canada.

McCraae, R. R., and John, O.P. (1992). An Introduction to the Five-Factor Model and its Applications. *Journal of Personality*, 60: 175–215.

Mori, J., Prendinger, H., and Ishizuka, M. (2003). Evaluation of an Embodied Conversational Agent with Affective Behavior. In *Proc. of the AAMAS03 Workshop on Embodied Conversational Characters as Individuals*, Melbourne, Australia.

Moundridou, M. and Virvou, M. (2002). Evaluating the Persona Effect of an Interface Agent in an Intelligent Tutoring System. *Journal of Computer Assisted Learning*, 18(3): 253–261.

Nass, C., Isbister, K., and Lee, E. J. (2000). Truth is Beauty: Researching Embodied Conversational Agents. In Cassell, J., Sullivan, J., Prevost, J., Churchill, E., editors, *Embodied Converastional Agents*, pp. 374–401, MIT Press, MA, USA.

Nass, C. and Lee, K. M. (2000). Does Computer-Generated Speech Manifest Personality? An Experimental Test of Similarity-Attraction. In *Proc. of CHI 2000*, pp. 329-336. The Hague, The Netherlands.

Nielsen, J. (1993). *Usability Engineering.* Morgan Kaufmann, San Francisco.

Paradiso, A. and L'Abbate, M.A. (2001). A Model for the Generation and Combination of Emotional Expressions. In *Proc. of the AAMAS01 Workshop on Representing, Annotating and Evaluating Non-Verbal and Verbal Communicative Acts to Achieve Contextual Embodied Agents.* Montreal, Canada.

Pelachaud, C., Carofiglio, V., De Carolis, B., de Rosis, F., and Poggi, I. (2002). Embodied Contextual Agent in Information Delivering Application. In *Proc. of First International Joint Conference on Autonomous Agents and Multi-Agent Systems*, Bologna, Italy.

Pelachaud, C. and Poggi, I. (2002). Subtleties of Facial Expressions in Embodied Agents. *Journal of Visualization and Computer Animation*, 13: 301–312.

Prendinger, H., and Ishizuka, M. (2002). Social Role Awareness in Animated Agents. In *Proc. of AAMAS02*, pp. 270–277.

Prendinger, H. and Ishizuka, M. (2003). Designing and Evaluating Animated Agents as Social Actors. *IEICE Transactions on Information and Systems*. E86-D(8):1378–1385.

Reeves, B. and Nass, C. (1996). *The Media Equation — How People Treat Computers, Television and New Media Like Real People and Places*. Cambridge University Press, Cambridge.

Resnick, P.V. and Lammers, H.B. (1985). The Influence of Self-esteem on Cognitive Responses to Machine-Like Versus Human-Like Computer Feedback. *The Journal of Social Psychology*, 125: 761–769.

Rickenberg, R. and Reeves, B. (2000). The Effects of Animated Characters on Anxiety, Task Performance, and Evaluations of User Interfaces. In *Proc. of CHI 2000*, pp. 49–56. The Hague, The Netherlands.

Russell, J.A. and Mehrabian, A. (1977). Evidence for a Three-Factor Theory of Emotions. *Journal of Research in Personality*, 11: 273–294.

Sanders, G. and Scholtz, J. (2000). Measurement and Evaluation of Embodied Conversational Agents. In Cassell, J., Sullivan, J., Prevost, J., Churchill, E., editors, *Embodied Converastional Agents*, MIT Press, pp. 347–373.

Sproull, L., Subramani, M., Kiesler, S., Walker, J.H., and Waters, K. (1996). When the Interface is a Face. *Human-Computer Interaction*, 11: 97–124.

Van Mulken, S., André, E., Müller, J. (1998). The Persona Effect: How substantial is it?. In *Proc. of HCI1998*, pp. 53–66. Sheffield, UK.

Witkowski, M., Arafa, Y., and De Bruijn, O. (2001). Evaluating User Reaction to Character Agent Mediated Displays Using Eye-Tracking Technology. In *Proc. of the Workshop on Information Agents in E-commerce; Agents and Cognition*, AISB Convention, York, UK.

Xiao, J., Stasko, J., and Catrambone, R. (2002). Embodied Conversational Agents as a UI Paradigm: A Framework for Evaluation. In *Proc. of the AAMAS02 Workshop on Embodied Conversational Agents: Lets Specify and Compare Them!*, Bologna, Italy.