

Evaluating Affective Conversational Agents: questionnaire analysis, behavior observation or both?

Irene Mazzotta, Giuseppe Clarizio, Fiorella de Rosis and Enzo Silvestri

Intelligent Interfaces, Department of Informatics, University of Bari

<http://www.di.uniba.it/intint/>

We wish to design an Embodied Conversational Agent whose role is to promote health-related behaviors. After a theoretical work that we performed in cooperation with cognitive psychologists (Miceli and Poggi) in the scope of WP8, we hypothesise that rational and emotional persuasion strategies may be successfully integrated to reach such a goal; we wish to test whether this is true, and how the two forms of persuasion may be blended, to produce effective persuasion dialogs.

From our previous experiences of ECA's evaluation studies, in the scope of a previous European Project, we could experiment the limits of 'classical' questionnaire studies in this domain. First of all, limited involvement of users in the definition of what the ECA should do and how it should interact with the users; but, in addition, strong influence of the agent's expressions and degree of realism on the subject's evaluation of the dialog *content*. We were convinced that filtering evaluation from this influence is essential, considering that artificial characters are still in a early stage of development, in spite of the considerable research efforts spent in the last ten years in this area.

Our evaluation procedure was based on the principle of *iterative design*, with integration of prototyping with evaluation.

In a first step, we performed an electronic brainstorming: we built a website to collect the viewpoints of subjects with various backgrounds (from 'experts in persuasion' to 'not expert in the domain'). We analysed their proposals, to sketch some hypotheses about 'natural' persuasion strategies which are suggested by humans, to interate with the results of our theories.

In a second step, we compared alternative persuasion strategies: we selected four 'typical' strategies, that we simulated in 'virtual dialogs' between two Embodied Agents, again with a website; we compared the effectiveness of these strategies with an open questionnaire in which 'evaluations' were integrated with 'suggestions';

The third step was aimed at implementing and evaluating direct interaction of users with our ECA: we performed a multistep Wizard of Oz study, to observe the users' behavior when interacting with the ECA in the mentioned domain and collect their subjective evaluation. At every step, we revised the agent's conversational attitude, according to the results of the previous step. The following is a short reflection on this multistep evaluation approach.

Electronic brainstorming produced a wide repertory of persuasion strategies, some of which we had not imagined (much larger use of ‘appeal to emotions’ than expected), and therefore a corpus to compare with our previous hypotheses and with which to enrich our theories. But also a new question: are these strategies really considered as ‘persuasive’ by human users?

Evaluation of alternative persuasion strategies was performed by asking subjects to ‘witness’ virtual dialogs between two ECAs. In these short dialogs, Alice tries to persuade John to eat vegetables, and the subjects are asked to put themselves ‘in the shoes of John’. A final *open* questionnaire asks them to evaluate the characters’ performance and the dialog content and to justify their evaluations.

Evaluation of the user behavior when interacting with the ECA was performed with WoZ studies: we designed and implemented a domain-independent tool to simulate dialogs with ECAs; this tool enables subjective and ‘natural’ (icon-based) evaluation of individual agent moves (not compulsory) and final, ‘subjective’ and compulsory evaluation of the dialog and the agent (a questionnaire with Likert scales etc...), in addition to a collection of logs of dialogs. The logs of dialogs may be employed to analyse the subjects’ attitude as an indirect evaluation of their persuasion level and their ‘interpersonal stance’ with the ECA. The following measures of the subject’s attitude can be performed (for instance):

- a) *level of involvement in the dialog*, as a function of the *dialog duration* and of the *average length* (in characters) *of the user moves*;
- b) *degree of initiative in the dialog*, as a function of the *percentage of questions raised by the subject* over all the dialog moves,
- c) *interpersonal stance* (from parsing of the moves), through language signs like: friendly self-introduction, self-disclosure, personal questions to the agent, irony, familiar and dialectal language, etc.

We see the following advantages of WoZ studies:

- opportunity to *integrate ‘subjective’ and ‘objective’ data, which contributes to evaluate the system in depth*;
- opportunity to *evaluate the dialog not merely from the sum of evaluations of individual moves and not only from the ‘degree of expressiveness and realism’ of the ECA*: in our experiments, subjects evaluated the individual moves quite unfrequently, as these seemed to ‘distract’ them from the dialog; the agent’s expression did not seem to be the main concern of users (while it tends to be so in questionnaire evaluation of monologs!);
- *subjective evaluations and observation of the subject behavior provide different kinds of information*; their results are (apparently!) inconsistent; on the contrary, they contribute to illustrate the problem under study more in depth.

Why an iterative model for the WoZ studies? Because it enables implementing the typical ‘iterative design’ approach to system implementation. At every step, the ‘agent moves’ are revised (in content, length, style,

...) and new moves are added, to solve the problems found in the previous step. In particular, the kind 'empathic attitude' the ECA should show, to favour the user-ECA relationship and the persuasion strength, is assessed.

What has to be done, to make a WoZ study?

- Formulate a *scenario* which describes the goal of the dialog
- Sketch a *dialog plan* that the wizard will be requested to follow
- Prepare a *list of agent moves* for every dialog step;
- *Categorize agent moves* according to some criteria;
- Prepare a *final questionnaire*

In the Workshop, in addition to enable the participants to test and evaluate in depth our tools, we plan to discuss the main findings of our iterative and integrated approach to evaluation: rather testing a lab hypothesis, we could enrich our theory-based hypotheses with proposals from potential users; we could filter (at least in part) evaluation of the message content from the effect of the agent's appearance; finally, we could observe the effects of our intended dialog on potential users and revise it until approaching the desired effect.