



**Österreichisches Forschungsinstitut für /
Austrian Research Institute for /
Artificial Intelligence**

TR-2011-25

*Dirk Heylen, Rieks op den Akker, Mark ter Maat,
Paolo Petta, Stefan Rank,
Dennis Reidsma, Job Zwiets*

**On the Nature of Engineering Social
Artificial Companions**

- Freyung 6/6 • A-1010 Vienna • Austria •
- Phone: +43-1-5336112 •
- <mailto:sec@ofai.at> •
- <http://www.ofai.at/> •



**Österreichisches Forschungsinstitut für /
Austrian Research Institute for /
Artificial Intelligence**

TR–2011–25

*Dirk Heylen, Rieks op den Akker, Mark ter Maat,
Paolo Petta, Stefan Rank,
Dennis Reidsma, Job Zwiers*

**On the Nature of Engineering Social
Artificial Companions**

The Austrian Research Institute for Artificial Intelligence is supported by the
Austrian Federal Ministry for Science and Research and the
Austrian Federal Ministry for Transport, Innovation and Technology.

On the Nature of Engineering Social Artificial Companions¹

Dirk Heylen, Rieks op den Akker, Mark ter Maat, Job Zwiers

Abstract:

The literature on social agents has put forward a number of requirements that social agents need to fulfill. In this paper we analyze the kinds of reasons and motivations that lie behind the statement of these requirements. In a second part of the paper we look at how one can go about engineering the social agents. We introduce a general language in which to express dialogue rules and some tools that support the development of dialogue systems.

1 Introduction

The past decade has shown a clear growth in interest to study long-term relationships between humans and artificial companions such as agents and robots, witness the current surge in EU projects such as Companions (Smith et al., 2010), LIREC (Castellano et al., 2008), and SERA. An important theme in the study of such agents is the social relationship that gets constituted in the interaction and that should lead to a prolonged wish to interact with the artificial companions. From an engineering point of view the question is "How can we design and construct such artificial social companions?" But also, "What does it mean to be social for an artificial companion?" "Is it the same thing as for a real, human companion or can artificial companions be social in a different way?" Following this question is another one: "What should artificial companions be capable of doing?" And then there is the ultimate engineering question: "How can we get the artificial companions to behave in this way?"

This paper has two parts. In the first part we look at the issue of requirements. Several studies concerned with social companions have looked at what skills an artificial companion should be able to display. In this paper we are not so much concerned with the actual requirements that are brought up in the literature, but we identify the ways in which the requirements were derived; their basis; the way their introduction is motivated. The question here in this part is: "How does *requirement engineering* in this field work?" (rather than "What are the requirements on social agents?"). In the second part we will be concerned with an exercise in engineering a particular social skill and reflect on the nature of *engineering social skills* in general.

The conclusion that we will draw from this is the following. In order to build artificial social companions, we need a thorough understanding of the dialectic interaction between the human and the machine. One aspect of this is an understanding of what stance the human takes towards the technology and the other is what the technology assumes about the human in interaction. To arrive at an understanding of how artificial companions work it is therefore necessary to study real interactions between humans and machines in ecologically valid situations.

¹ This is a preprint of an article submitted for consideration in the Journal Applied Artificial Intelligence 2011 [copyright Taylor & Francis]; Applied Artificial Intelligence is available online at: <http://www.tandf.co.uk/journals/journal.asp?issn=0883-9514&linktype=44>.

2 Requirements Engineering for Artificial Companions

As a first step in designing an artificial social companion, one might ask the question what such a thing should be capable of doing or to look for guidelines to design them. These are questions that are part of requirements engineering. The literature on requirements engineering provides many methods to elicit requirements from potential users and other stakeholders. In this section we will not review those, but we will look at how requirements, or design features, of an artificial companion are motivated in the literature that deals with building or evaluating them. We will distinguish various types of studies, embodying different methodological choices and therefore also different kinds of arguments for having or not having certain design features. They are amply illustrated in the following sections.

2.1 Lab Studies

A first way to study what an artificial companion should be capable of or what it should look like involves creating mock-up or real companions and ask people what they think of it in a lab situation. Here are some examples of such studies.

Studies

Goetz et al. (2003) tested two contrasting social psychology hypotheses: The “positivity hypothesis” asserts that attractiveness, extraversion and cheerfulness correlate generally with acceptance and compliance by human peers, while the “matching hypothesis” states that appearance and social behaviour of a robot should be matched to the seriousness of the task and situation. In the study people were presented with pictures of robots that differed along three dimensions: age, sex, and human versus mechanic looking. Participants were asked which robot appearance they would prefer for different tasks. For female-looking robots, participants preferred the humanlike robots to the machine-like robots for most jobs: actress and drawing instructor (artistic jobs), retail clerk and sales representative (enterprise jobs), office clerk and hospital message and food carrier (conventional jobs), aerobics instructor and museum tour guide (social jobs). Machinelike robots were preferred over the humanlike robots for jobs including lab assistant and customs inspector (investigative) and for soldier and security guard (realistic). The patterns for the masculine looking robots were not as strong but went generally in the same direction. In this case participants slightly preferred humanlike robots for artistic and social job types, but preferred machinelike robots for realistic and conventional jobs. It was concluded that a robot’s appearance and behaviour indeed do influence people’s perceptions of the robot’s capabilities and social behaviour skills and their willingness to listen to the robot’s instructions. Furthermore, elicited responses were framed by people’s expectations of the robot’s role in the situation. Thus, it was the more differentiated matching hypothesis that found support, e.g. in terms of more humanlike, attractive, or playful robots not necessarily being considered to be more compelling, in particular when such appearance did not match the given task context.

A similar experiment was carried out by Hegel et al. (2009). They looked at the influence of visual appearance of social robots on judgments about their potential applications. They chose videos of twelve robots, Barthoc Jr., iCat, AIBO, BIRON, KeepOn, Kismet, Leonardo, Robovie, Repliee Q2, ASIMO, Paro, and Pearl and thirteen “jobs”: security, research, healthcare, personal assistant, toy, business, pet, entertainment, teaching, transport, companion, caregiver, public assistant. Among the results of the study was that the application fields entertainment, pet, toy, and companion strongly correlate with an animal-like appearance, while the others strongly correlate with human-like and functional-like appearances. The visual appearance of robots was shown to be a significant predictor for the appreciation of applications. Additionally, the ratings showed an attractiveness bias, in that robots that were judged as more attractive were also evaluated more positively on a liking scale.

Several studies were conducted to determine whether a robot or agent interface would be preferred over other types of interfaces such as a simple text interface, for instance. Looije et al. (2006) and Looije et al. (to appear) showed that both children and older adults preferred to interact with an animated iCat robot over a simple text interface in studies on support of diabetes treatment. A pilot study by Tapus et al. (2009) with a socially assistive robot for elderly people with cognitive impairments showed that the robot’s physical embodiment played a key role in its effectiveness. They compared an embodied system with a computer interface in the context of a music-based cognitive game. Not only did users prefer the embodied system, but the embodied system also had an overall effect on sustaining and improving the performance on the task.

Walters et al. (2008) studied robot appearances as well. They did an experiment with three different kinds of robots: a mechanoid robot, a basic (humanoid) robot and a humanoid robot². Students did not interact with the robots themselves, but expressed their opinions about the robots in a video-based human-robot interaction. The results showed that overall participants tended to prefer robots with more human-like appearances and attributes; however it was also found that these preferences seemed to differ with the students’ personalities: introverts and participants with lower emotional stability tended to prefer the mechanical appearance to a greater degree than other participants.

The Merits and Limitations of Lab studies

Studies like the ones described above follow a classical experimental design with controlled variables. In several of the studies, the participants did not interact with an interactive system, but were shown pictures or videos. In other studies mock-up of interactive systems were constructed. Through questionnaires they were asked to rate their likings and dislikings. Studies such as these are important for the design of artificial companions mainly because they can point out the many factors that play a role in the appreciation of a particular design feature.

² In contrast to the humanoid robot equipped with two seven-degrees of freedom arms capable of human-like gestures (e.g. waving), the basic robot was equipped with just one single-degree of freedom arm which enabled it to make pointing gestures. The mechanoid robot was equipped with a simple single-degree of freedom gripper that could only be moved up or down.

However, they are limited by the fact that they consider only a limited number of design cases that were laid down by the experimenter. In contrast, participatory design studies, might have used a constructive approach in which users could have co-shaped the outlook and the interaction skills of the companion.

The studies involve interactions with systems out of the context for which they are ultimately intended, often involving participants that are not real users, of the system but participants in an experiment.

Although the findings may thus not generalize to other cases, or to actual (long-term) use of a system, they do provide potential insights into tendencies related to preferences and the many factors that influence these likings which need to be taken into consideration in actual designs. When the studies are used to engineer and design they need to be deliberated carefully. To what extent is the system to be designed comparable to the system tested in the experimental situation? To what extent are the participants similar and in what ways to they differ? To what extent is the situation or the context of use and the application the same and in what ways to they differ? The results of experimental studies like the ones above can guide in two ways the engineering of artificial companions. One is in choosing particular design features based on their findings. The second is in inspiring the set-up of pilot studies for the application to be developed.

2.2 Extrapolation from Human-Human Interaction

In the current thinking about artificial companions, there seem to be two strands. One way to go is to develop pet-like artifacts such as Paro and the other is to develop human-like artificial companions. In the second case, it seems to be rather obvious that design guidelines for artificial companions follow the model of the human companion. Whatever makes the human companion work, whatever behaviours and skills the human behaviour display, the artificial companion should be capable of those as well.

Generally it is argued that it is desirable for virtual humans and humanoid robots to be "*life-like*" (Predinger and Ishizuka, 2004). Various kinds of arguments may be given for this point of view. First, it is often argued that agents should resemble people because they perform roles and fulfil tasks that are normally performed by humans. Second, users are used to and prefer to interact with other humans. However, it has also been claimed that social robots should not resemble humans too much. When robots look like humans this may elicit strong expectations about the robot's social and cognitive competence. If such expectations are not met, then the user is likely to experience confusion, frustration and disappointment. For a discussion of this effect that is known as the "uncanny valley" we refer to Dautenhahn (2002) who argues that this effect is highly context dependent, though. In some situations it is more acceptable that expectations are not met than in other situations. For example, when a robot serves as a servant for elderly people it is less acceptable when expectations are not met than in situations where robots serve as a kind of toy (e.g. AIBO).

On the basis that artificial (human-like) companions should be able to display behaviours and skills similar to human companions, Fong et al. (2003) listed the following six types of

requirements which are often cited in the literature. Social robots and agents should be able to:

1. Express and/or perceive emotions
2. Communicate with high-level dialogue
3. Learn/recognize models of other agents
4. Use natural cues (gaze, gestures, etc.)
5. Exhibit distinctive personality and character
6. If possible, learn/develop social competencies

Some form of communication is needed as a basis to establish a social relation (Castellano et al. (2008); Duffy (2003), Fong et al. (2003); Green et al. (2004), Kidd (2008); Li et al. (2006)). In embodied conversational agents this is often interpreted as a requirement for social robots and agents to possess conversational skills similar to humans, including ways to open and close conversations and showing engagement (Sidner et al. 2004); turn-taking; providing feedback (Kidd, 2008), contrast, and emphasis; showing attention; addressing (Bruce 2002); formulating sentences, and constructing multimodal communicative actions (see Cassell et al. 1999 for an overview). On the basis of such communication skills, further social interaction skills can be built. For instance, providing feedback can be important to motivate users or to express empathy (Blanson-Henkemans et al. 2009; see also below). Skills like these are often assumed to be important in human-robot interaction just because they are fundamental in human-human interaction. However, they apply in particular to humanoid robots and agents with dialogue abilities. For the case of zoomorphic agents/robots the case might be different (consider, for instance, the examples of Paro or AIBO).

Several authors focus on the importance of using “natural” cues in interaction – the verbal and nonverbal cues humans use effortlessly in interaction. They have studied the use of particular mainly nonverbal behaviours and their effects on the interaction and the social relationship: eye contact, look-at behaviours, head, arm, and hand gestures (see e.g. the references in the previous paragraph). Subtle gestures were found to have a positive effect on understanding what a robot was doing (Breazeal et al. 2005). One of the factors that Kidd (2008, p. 162) mentions as contributing significantly to the creation of a successful system is the appearance of eye contact and the movement of the eyes: “Numerous participants who had the robot (and countless others who have seen or interacted with it for short durations) noted during the final interview that the eyes of the robot and the fact that it looked at them drew them into interactions with the system and made it feel more lifelike. Users were clearly more engaged with a system that looked at them [...]” Also Kozima et al. (2003) point out the role of eye contact on joint attention as a prerequisite for social interaction. Bruce et al. (2002) indicate attention of the robot by having the robot turn to a person when it wants to address that person (see also the work by Sidner et al. (2004, 2005) on engagement).

It is interesting to consider the question though what the importance of such natural cues is exactly. Several conversational functions, for instance, can also be expressed by cues that are less natural. Kobayashi et al. (2008) used a blinking LED to notify a user about the robot’s internal state such as processing or being busy. The use of such a visual device made conversations go smoother and human users would not repeat themselves as often (see also Sengers 1999). As humans are – by nature – capable of devising new – and artificial (?) – ways of communication, perhaps a naturally artificial companion should be able to do the

same: develop an interspecies communication protocol instead of merely adapting to the human case?

The extrapolation of findings from human-human studies to human-agent interactions can be based on the idea that humans have a particular way of interacting that comes natural to them so that a system that exploits this type of interaction is also easier to learn and experienced as more pleasant. It can also be based on the goal of building agents or robots that are as much human-like as possible. The validity of extrapolating guidelines for the design of naturally interacting artificial systems from what we know of human-human interactions is often argued for by the Media Equation, which was demonstrated in a series of studies by Reeves and Nass and students (Reeves and Nass, 1996). Many of these first studies just used “old-fashioned” text-based interfaces, but since then evidence has been steadily extended in support of the broader claim that in interaction with computers, humans show the same relational stance towards the computer as they show towards humans. For instance, studies on this so-called CASA (“Computers Are Social Actors”) paradigm have shown that computers that use flattery, or which praise rather than criticize their users are liked better (see also Fogg and Nass, 1997; Johnson et al., 2004). Also, users prefer computers that match them in personality (Johnson and Gardner, 2007), a phenomenon known as the “similarity attraction” principle in social psychology (Byrne et al., 1986).

The premise that robots and agents should be human-like seems to be the dominant opinion of the field. However, as Kobayashi et al. (2008)’s successful example shows, this should not be mistaken for an immutable principle. When human-likeness is the goal, analysing human-human interactions in the selected domain and trying to copy the relevant behaviours on the system is a good way to proceed but care must be taken to identify the full scope of the related required capabilities and to assess the viability of their technical realisation.

Taking the human-human case as model seems an obvious way to go. It might be the goal that one sets oneself to achieve. However, one might also wonder whether artificial humans should be like the real ones in every aspect of their construction. Perhaps there is also an appeal in humans to engage in the exploration of other-hood that can be exploited in the design of artificial companions.

From an engineering perspective the choice for making artificial companions as much human-like as possible opens up a wide literature on the study of human-human interactions that can be taken as a basis for requirements engineering. It also offers methodological guidelines on how to proceed: look at human-human interactions first, analyse them and try to copy them. This was the methodological approach chosen in Cassell’s Gestures and Narrative Language Lab at MIT and is illustrated by the work of Bickmore (Bickmore 2003). However, it also leaves some questions unanswered. One is the engineering question: knowing what kind of thing to build, the question remains is how to build it. The second is: how do people actually and really interact with artificial companions?

The second question is a fundamental one and requires a third type of study in which actual artificial companions are put in the homes and the lives of real people. The question here should not just be to test how well the system that is put there works, but instead and

rather, how people in more-or-less everyday situations (ecological validity) deal with so-called artificial companions.

3 Modelling Social Interaction

In the SERA project, a fairly simple but robust artificial companion was put in real people's (not psychology student's) home (not the lab) to see how, over a course of ten days, the participants in the study would interact with such an artificial companion in their home. The interactions with the artificial companion – embodied through the Nabaztag³ – consisted of a number of pre-scripted dialogues that would be activated at certain times of the day, depending on a number of circumstances. The Nabaztag used speech to convey information. The participants had a series of cards with RFID tags to respond. The participants in the study were somewhat older people that tried to maintain an active lifestyle and so they would make plans to do a number of activities each week that would improve their health. The SERA companion knew about their planned activities and their goals and would inquire about them and offer advice. The interactions between the SERA companion and the participants were triggered by particular events. For instance, when the participants walked past the SERA companion the first time during the day, the companion would start a pre-scripted dialogue about their plans for the day. The SERA companion was 1) aware about the time of the day, 2) the activity plan of the participant of the day and 3) whether or not there was somebody in front of it. The SERA companion had a number of dialogues that it could perform depending on the situation. Next we present a commented version of one of the interactions to show the nature and complexity (or rather simplicity) of the dialogues and the problems that arise when the companion gets it wrong.

3.1 An example case

P4 (= participant 4) approaches the companion⁴. Approaching the rabbit, this sets off a PIR (passive infrared) sensor informing the system that someone is there. The first time in the morning that participants approach the robot, they are told what goals they have set in their activity plan (the activities they have scheduled for that day). In this case it is not the first time that P4 has walked past the SERA companion that day, but as she has not previously responded to the initial question, the companion asks her again⁵:

N1: Good morning, how are you?

<P4 shows the 60 minutes card>⁶

P4 responds by showing the SERA companion the "60" minute card (a card that is equipped with an RFID tag, so the companion can identify its content), because she wants to add 60 minutes of activity to the activity plan. However, showing such a card in this particular dialogue does not make sense to the companion. It expects a card that matches the question it asked. The companion therefore ignores the card and tries to complete its task of

³ <http://www.nabaztag.com/>

⁴ This analysis is a modified version of one provided by Sarah Creer.

⁵ We introduce each utterance with either N (=Nabaztag) or P (=participant) + a number identifying each utterance.

⁶ Between angle brackets we briefly describe actions that the participant takes that we can see from the video.

presenting the information on what it knows about the activity plan to the user. The companion responds with:

N2: If you wanted to tell me about your exercise then please can you remind me about that once I've just quickly told you about the weather first. <P4 shows 60 minute card> If it's okay to record you now, please could you press the video button? <P4 presses video button> Would you like to hear what the weather forecast is today?

While the companion is saying this, the user tries to interrupt it using the 60 minutes card again. However, the companion does not know how to deal with these interruptions and proceeds with its script. In response to the question about the video button, the participant presses the button, acknowledging her consent to the filming. The companion acknowledges this in return.

N3: video recording on.

In response to the second time that the user has shown the companion the 60 minute card, it still sticks to its script which is to ask the participant whether it is ok to inform her about the weather forecast, ask her about weighing herself and reminding her of her activity plan for the day. In the process of trying to convey that information to the participant, the participant seems only to hear that the companion has understood what it is that she wants to do, interrupting the companion by showing the companion and the camera the 60 minutes card. She is also ready to use the card after the reminder but then the companion tells her the weather as she is getting more frustrated.

N4: If you wanted to add some exercise to your log,

P1: Yep, one hour with my dog this morning.

N5: then please can I just first quickly tell you the weather and pass on how much exercise you've got planned today. If you could remind me after about adding exercise to your log, I'll do that for you..

P2: Okay

N6: The weather today is Mostly dry and sunny. Showers on hills. Cold again tonight.

P3: Thank you.

P4 is again seemingly trying to use the card again to input the amount of exercise that she has done but the companion continues to go through its morning routine. At this point, it is asking her about weighing herself, as a way of tracking progress in her healthy living schedule⁷.

N7: Some people find it useful to weigh themselves daily [as part of their exercise routine.]

P4: [I know. I don't want that reminder.]

N8: I hope you don't mind me asking but will you be weighing yourself today?

P5: How?

Due to a lack of response by P4, only then the dialogue is programmed to continue to provide the participant with their goal for the day. If the amount of exercise is over 30 minutes (as in this case) the user is given the NHS (National Health Service)

⁷ The square brackets indicate overlap between the utterances. In this case "as part of their exercise routine" overlaps with the utterance P4 (4th utterance by P4, participant 4).

recommendations for an older adult. There is a lot of overlap in this part of the interaction with the user trying to work out how to use the cards to respond to the questions.

N9: The [NHS recommends doing at least 30 minutes of activity 5 times a week.]

P6: [How do I tell you no? Will I weigh myself today? How do I say no?]

P7: [Is it not very nice face?]

<P4 shows frown face to companion>

N10: [From your plan it looks like you've got] a total of 140 minutes of planned activity today.

P4 interrupts this. She seems to still want to answer the weighing question and is looking to the cards to find the way to respond.

P8: [I'm not going to weigh myself today.] Okay, yeah, I know.

At the end of every sub-dialogue is an "ending" message which signals that the companion has optionally ended that section of dialogue. The participant can carry on dictating the direction of the conversation if they use any of the topic or minutes cards.

N11: Unless you'd like to talk about [anything else],

P9: [I want to record my]

N12: I'll talk to you later. Bye

P10: Don't go away. I want to record my activity companion. <P4 shows 60 minutes card to companion> And now you're not reading it. No, you are not reading it.

After P4 has shown the 60 minutes card again to the companion, it is now possible to change topic. It therefore asks if P4 would like to add something to her log. The participant has started to think that the companion has not read the input and starts to explain to it what she has done that morning even if it looks like it is struggling to store the numbers.

P11: My dog and I have been out this morning to the park.

N13: Would you like me to add something to your [exercise log?]

P12: [Yes.] 60 minutes.

P13: 60.

P14: minutes.

The companion has not been able to read the cards and instead uses its excuses to explain that the user needs to try again. The companions stays in the same state at this point but uses what is called "encouragements" which try and tell the participant that it is not their fault that the system is struggling, it is the system's own fault. The encouragements also give a bit of advice on where to place the tag so that it can be read by the Nabaztag.

N14: Sorry, it's sometimes a bit difficult for me to read the [cards, it's] my fault,

P15: [I know, difficult.]

N15: can you please try again?

P4 at this point is frustrated by the cards not working so, as she is explaining to the companion, she is trying to get the information across in a different way.

P16: Okay, now try two thirties.

N16: Please [can you try with the tag right against my nose?]

At this point it looks like P4 has given up on trying to interact with the companion and input her activity into the log and leaves after making her feelings clear to the companion.

P17: [No, no. I'm leaving. I've had enough]. You've not recorded, you've given me a weather report I didn't want. You have asked me to remind you about several things and then given me no way of doing it.

P18: I'm out of here.

<P4 leaves>

The companion at this point still assumes that the participant is trying to input information and so continues to give advice but does not have the information that the participant has left.

N17: Sorry, it's sometimes a bit difficult for me to read the cards, it's my fault, can you please try again?

This exemplary dialogue shows a typical way of humans interacting with a technology whose inner workings they only partially understand on the one hand, and of technology that is programmed and preset in a particular way as it was assumed and envisioned by the designers of the dialogue that the ritual interaction between a participant and the technology should work out.

3.2 Discussion

The dialogue shows that P4 and the companion have a hard time in aligning their individual intentions. Where P4 is determined to use the companion for administrating her activity, the companion has a different plan. The companion blindly follows the morning ritual. In the comments we see that there are good reasons for this: it is the companion's job to do so. Companion's utterance N2 in response to P4's showing the 60 minutes card signifies that the companion is not insensitive for this act of the participant. Moreover, the companion's interpretation of P4's act seems to match her intention. The companion seems to understand what P4 wants, but it seems also quite convinced of its own plan. The participant may expect that the companion itself will remember P4's intention. But N11 reveals that it doesn't really comply with her desire.

N11: Unless you'd like to talk about [anything else],

P9: [I want to record my]

N12: I'll talk to you later. Bye

P10: Don't go away. I want to record my activity companion.

The dialogue shows that P4 has problems accepting that the companion follows strict procedures. We can only guess whether she understands that the companion is just a "stupid" mechanical tool, and that there is no other way out for her than to follow the procedures that make the companion act as it does and hope that she will have her turn eventually.

Imputing interaction skills Sometimes P4 gives up pushing her own desires, for example in the dialog fragment from utterance N4 to P3. Overall the impression is that she sees the companion as something you can have a normal conversation with. As if she thinks the companion understands what she is saying. This includes the normal turn-taking behaviour, as we see for example in this sequence.

N4: If you wanted to add some exercise to your log,
P1: Yep, one hour with my dog this morning.
N5: then please can I just first quickly tell you the ...

The layout suggests that N5 is the companion's response to P1 (in the same way as P1 is a response to N4) but actually N4 and N5 are inseparable parts of one single turn that doesn't allow being interrupted. Also in this respect, regarding turn-taking and the sensitivity for interruptions, the perception that P4 has about the capabilities of the companion does not align well with the real affordances of the companion. One might wonder whether P4 really thinks that the companion understands her. There is a good chance that she doesn't and that she is just saying aloud what she thinks. (We expect she knows that the only way to interact with the companion is by showing one of the RFID cards and that talking to it is useless.) She doesn't really address the companion nor does she expect that it will understand her. Her awareness of the fact that her interaction with the companion is being recorded may influence her behaviour. Does she give a performance? Who is she addressing when she explains the companion why she isn't happy with the companion's behaviour? Does she address the companion when she accounts for her leaving a further conversation? Or does she address the designers, the ones behind the curtains that pull the puppet's strings?

Mental Model The image or "mental model" that P4 has of the companion, the imagined dialog partner she has a conversation with, is largely influenced by the suggestion that comes from the fact that the companion produces sounds that make the impression of a spoken utterance. It is only a small step then to the imagination that someone is talking to her, someone who belongs to her language community and that shares with her a wealth of common knowledge, someone who knows about the practice of having a conversation with all social rites and conventions that go with that practice. This is basically as automatically, as we see a meaning expressed in words we encounter. The dialogue shows that P4 needs to adjust this image.

At the same time we can try to make our social agents become more like the image suggested. Here we take the perspective of the designer of social agents. After all it is technology that makes the companion produce sounds so that she believes it talks. The design of an interactive system is based on an implicit or explicit model that describes what the possible actions are that the participant can perform in any state of an interaction with the system. According to good principles of HCI design the participant interface of the system should be designed so that at any time it makes clear to the participant what the affordances of the system are. What are the possible actions and their effects? The designer of the system not only programs the system how to act but thereby implicitly also programs the participant how she is expected to act. The interaction with the system works then as long and as far participant and system share the designed model and adhere to this shared model; not unlike a tool works as long as participant and system adhere to the manual that says how to use it.

Designing Naturalness Designing a system that works in an intuitive way for naive users is hard because the naive user does not think in terms of well-defined actions performed at well-defined moments during an interaction. The naive users simply act according to their

natural way of doing things. They don't select and use words for example because they know that these are the words that do the job, they just say what they think.

The design of a system that works in an intuitive way works out well if and as far as the designer has succeeded to model the way the naïve user works (and talks), so that the effects of the words chosen by the user as input for the system conform to the meaning that the user intended when she uses these words. The words act in the system according to the meaning and intention the user had with them.

```
N4: If you wanted to add some exercise to your log,  
P1: Yep, one hour with my dog this morning.  
N5: then please can I just first quickly tell you the ...
```

The interaction N4, P1, N5 demonstrates how an implementation of the mechanical model of the naïve way of doing things as it could have been programmed in a dialog script aligns with the naïve and "normal" way that the user behaves. In fact, the alignment in the above fragment is by chance, not by design. As designers we know that the system would continue with N5 whatever the user said at P1.

Conclusion Many of the phenomena we have seen in the interaction show that the alignment between the companion and the participant is not as fine-grained as a naïve user – who is a practical expert in real-time face-to-face interaction – would expect. The fact that the companion keeps talking when P4 has already left the scene in disappointment witnesses the fact that it speaks in a void (like the participant in a telephone conference who kept on talking long after the conference was closed). How can we improve on this? There may not be a principal solution to the problem but we may be able to improve the interaction by reconsidering the possibilities to make the agent more sensitive for feedback while acting and act accordingly.

4 Potential Models

The dialogue above is paradigmatic for the type of phenomena that we have encountered in the recordings of the interactions between the robot and the human participant. They show the robot needs some more social skills in order to be an acceptable and useful partner for the participants. In particular, we have seen that

- the companion is in need of awareness of the state of attention of the human participant;
- the human participant needs to be able to interrupt the robot in the middle of an ongoing dialogue, even in the middle of a sentence;
- the companion needs more flexibility in handling communicative behaviour that was already planned and in execution;
- the human participant is in need of being able to correct the companion when it has made false assumptions, not only regarding his willingness to interact about a certain topic but also regarding the content of what is being said or believed by the robot.

What this particular example shows is that to realize the social skills, the system needs the abilities to simultaneously perform a number of interacting perception and action processes and to adjust its behaviour either in a reactive way (on the level of behaviour realisation) or

in a deliberate way, on the level of dialog planning. This will make the agent less “autistic”, more sensitive for interruptions and other actions by the user. This allows for a mixed initiative conversation (topic switch initiated by the human) and more natural turn-taking behaviour, a basic social competence.

We show how we can model what could be going on in these dialogues and how the models for the dialogue should be adopted accordingly. Below we provide the models that we came up with. The rules have been implemented in a simulation program that can execute the models. The purpose of presenting them in detail here is to show how a simple skill for humans (turn-taking, topic switching) becomes soon rather complex when it is modelled explicitly. Clearly, having such models is not a guaranteed success that the dialogues would work out correctly: they presume that the rabbit is able to perceive and understand the user actions correctly.

4.1 Global setup

The global setup of our system is to separate the agent model of the companion agent, hereafter called “the agent”, into a large number of small state machines. Some of these, like the “TurnTaking” model focus on turn taking behaviour such as getting the turn, giving the turn, interrupting etcetera, while ignoring the actual dialogue content, where others model the interaction about some specific topic. Some of the models are meant to *observe and interpret* ongoing behaviour. For example, the act of speaking has a different interpretation when the speaker is supposed to have the floor or not. In the latter case it is interpreted as an *interruption*, rather than a normal communicative act.

For the specification of the turn-taking behavior, the interruptions, topic switches and the topical dialogues we use Hierarchical State Machines. Hierarchical State Machines (“HSM machines”) are a form of finite state machines with some elements from Harel’s StateChart formalism. (See [Harel 1987]). Basically, this formalism extends well-known Finite State Automata by means of two mechanisms, both taken from the StateCharts model: 1) The possibility to group a number of states together in a *super state*, in a *hierarchical* way. 2) The notion of *parallel composition*, where two or more state machines operate in parallel. We will introduce the basic notions by means of some simple examples.

Consider the Simple HSM, called “Weather Dialog”, in Figure 1.

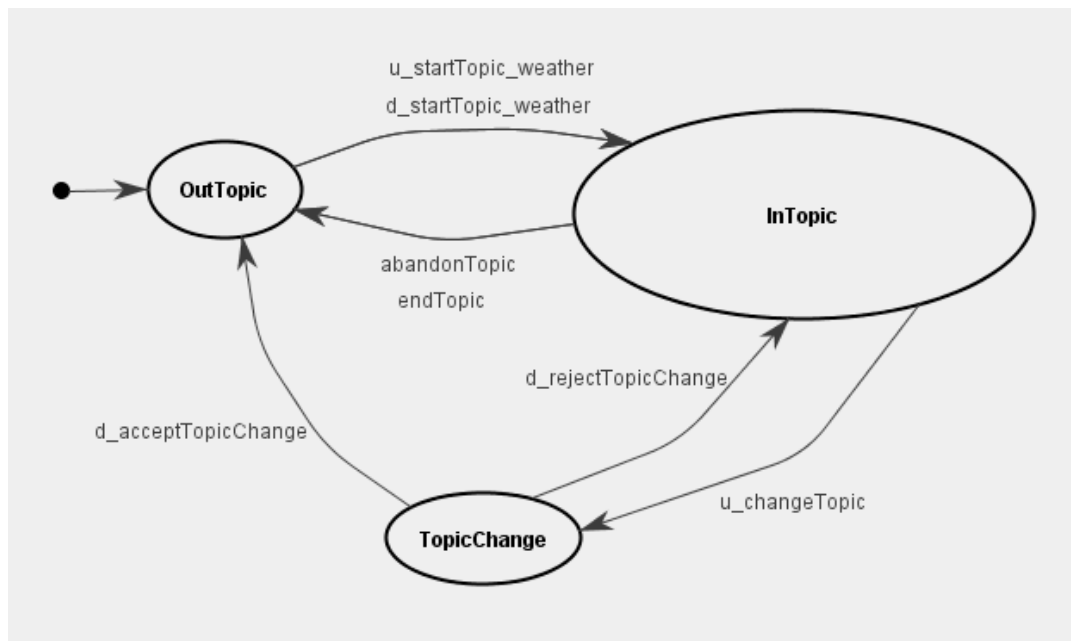


Figure 1 Simple HSM ("Weather Dialog")

This HSM is just a finite state machine with three *states*: "OutTopic", "InTopic", and "TopicChange". One state (the "OutTopic" state) is marked as being the initial state. Between states are *labeled transitions*, like for instance "u_startTopic_weather", informally denoting that the user stated that he or she wanted to discuss the weather. There is a similar transition labeled "d_startTopic_weather", denoting that the Nabztaq agent took the initiative to talk about the weather.

In general, the models from the SERA system have different types of labels on the arrows. The labels that have a "u_" prefix are observations that the agent does about the state of the user. The others are actions that are taken by the system when going from one state to another.

In the Simple HSM, it is *not* shown what exactly goes on while in the "InTopic" state. It is clear however, that the user could end (in a regular way) or abandon the ongoing discussion. As a third alternative, the user could propose to change topic, by means of a "u_changeTopic" transition, moving to the "TopicChange" intermediate state. From there, the rabbit agent could then choose to either accept or reject that change, moving either to the "OutTopic" state or back into the "InTopic" state. The set of all transition labels that occur within an HSM is called it *alphabet*. We will come back to it below, when we discuss parallelism.

Our next step is to *refine* this abstract view into a more concrete one, for instance, as in the diagram of Figure 2.

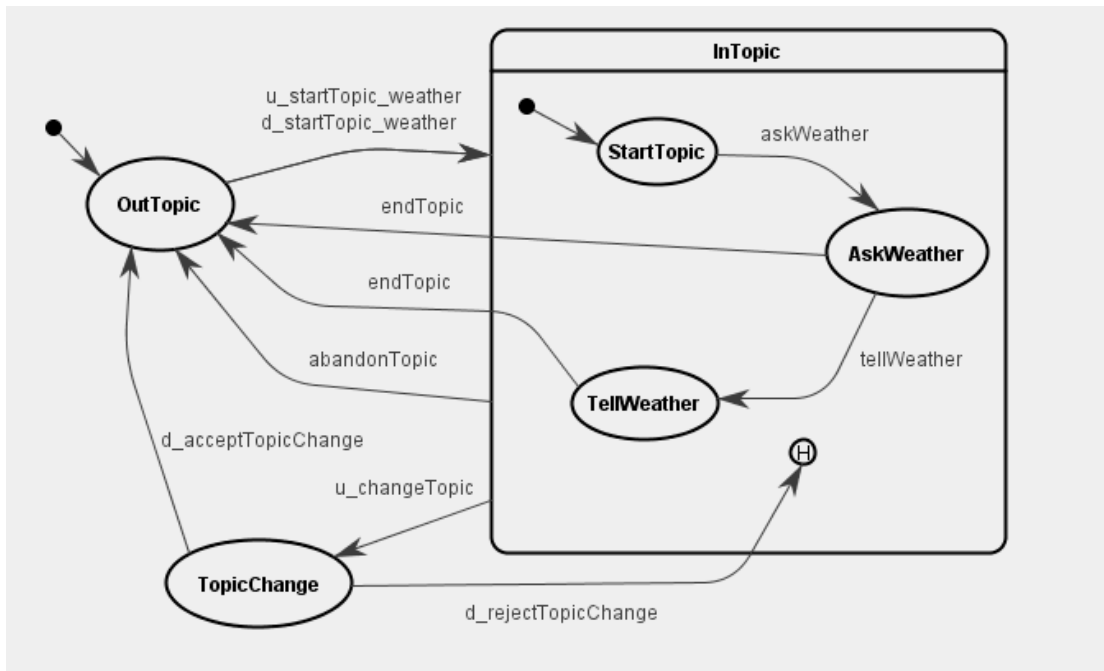


Figure 2 Illustrating sub states and super states

Here we see one major new element: the "InTopic" state, which was a simple state in the previous diagram, has now become a *super state*, containing three *sub states* "StartTopic", "AskWeather", and "TellWeather". The "u_startTopic_weather" and the "d_startTopic_weather" transitions from the "OutTopic" state to the "InTopic" state will actually result in a transition to the "StartTopic" sub state of "InTopic", since this sub state is designated as the initial state of the "InTopic" state. The simple dialogue then proceeds, first by asking whether the user actually want to be informed or not. If so, a transition to "TellWeather" is made, otherwise the dialogue is ended immediately. In both cases, the dialogue ends in a regular way by means of an "endTopic" transition, straight towards the "OutTopic" state.

An interesting new element is that there is also an "abandonTopic" transition, not from any sub state, but rather from the "InTopic" super state itself, as a whole. This denotes that when the user triggers such a transition then, regardless of the actual sub state in which the system currently resides, a transition to "OutTopic" will be made. In this tiny example we could easily have introduced three such "abandonTopic" transitions, one from every sub state. But one should really consider here some more complex discussion topic resulting in a super state with potentially dozens or hundreds of sub states. Then a single transition from the super state replaces a lot of similar transitions from sub states. Moreover, the design becomes more robust. For instance, consider that somewhere in the future we would replace the rather simple dialogue with a more elaborate one. Then we can simply work on more elaborate versions of "InTopic", and the "abandonTopic" transition will automatically be one of the possible transitions for any new sub states that we might introduce.

A final new element in the diagram is the "d_rejectTopicChange" transition towards a so called "history state", denoted by the H-labeled state. It denotes that a transition will be made to that particular "InTopic" sub state that the system was in at the moment we actually left "InTopic" when the "u_changeTopic" transition was made.

A second major concept in StateCharts is the idea of *parallel composition*. The idea is that two or more state machines are active simultaneously, each controlling one particular aspect of the systems overall behaviour. The main idea here is that each of these (parallel) components has its own internal state which might be quite independent from the states of other components. The way these state machines interact is by means of *joint transitions*. For an example, consider the state machine called "User Participation" in Figure 3.

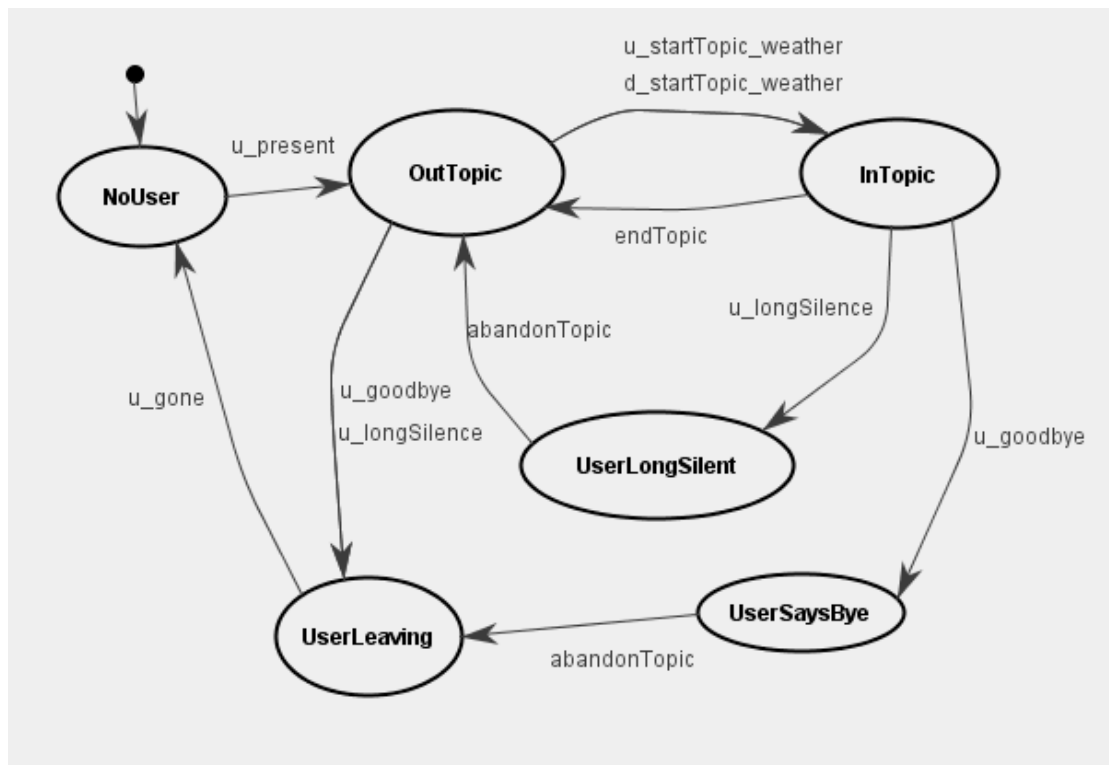


Figure 3 Parallel Composition in HSM ("User Participation")

The "UserParticipation HSM has certain transitions with labels identical to similar transitions in the "WeatherDialogue" HSM that we discussed above (Figure 1). For example, it contains transitions with labels like "u_startTopic_weather" or "endTopic", also part of the alphabet of the "WeatherDialogue" model.

The *rule* for parallel composition of HSM's is that whenever some label belongs to the alphabet of some HSM "H", then transitions with that label can occur only when the "H" machine can perform such a transition. The *effect* is that when some label occurs within the alphabet of two (or more) HSM's, then transitions for that label can only occur when *all* of these HSM's are ready to take the transition, and when this happens, a single joint transition occurs.

Returning to our example, we see that when "UserParticipation" and "WeatherDialogue" are simultaneously active, "running in parallel", then a transition like "u_startTopic_weather" can occur only when both HSM's are ready for taking it. And when they do, the transition will take place as a simultaneous joint transition. In our example this would imply the following: although our "WeatherDialogue" HSM is eager to start its weather dialogue right away, the "UserParticipation" will initially block the "u_startTopic_weather" (or "d_startTopic_weather") transitions, since these are not outgoing transitions from the "NoUser" state. In state machine parlance, these transitions are not *enabled*. Only *after* the "UserParticipation" HSM has taken its "u_present" transition, signifying that the presence of

some user has been detected, will the "startTopic_weather" transitions become enabled. One might ask why the user could trigger an "u_present" transition without cooperation from the "WeatherDialogue" HSM. The answer is that "u_present" is a label that does not belong to the alphabet of "WeatherDialogue", so it cannot be enabled or disabled by that HSM. Informally, one could say that an HSM puts constraints on transitions, but only on those transitions with labels belonging to its alphabet.

The "UserParticipation" HSM focuses (exclusively) on one particular aspect of user behaviour: whether the user is actually participating in any dialogue at all, or whether any user is present at all. For this aspect, we see, again, states labeled "OutTopic", "InTopic", with similar transitions as we saw before, for the abstract version of the weather dialogue diagram. Unlike that diagram, the "InTopic" state here remains as is, that is, it is *not* refined into sub states, since that is considered not relevant as far as *user participation* is concerned. Rather, our current diagram deals with aspects like, for instance, the difference between "ending" a topic the normal way, or "abandoning" it: The "endTopic" is enabled whenever we are in the "InTopic" state, but the "abandonTopic" is *not*. It will become enabled only after the "UserParticipation" HSM takes either a "u_longSilence" or a "u_goodbye" transition, signifying that a long silence on behalf of the user has been detected, or that the user actively said "good bye" to the system. In both cases the new state will then enable the "abandonTopic" transition while simultaneously disabling the "endTopic" transition. Compare this with the "WeatherDialogue", where both transitions are always enabled while in the "InTopic" super state. Apparently, for *that* particular HSM the reason for one or the other transition is of no concern.

This example, in a nutshell, shows one of the main reasons for modelling using parallel composition: it allows for separation of concerns, where each of (potentially many) specialized HSM machines deals with one and only one particular aspect of behaviour. Each of these HSM's can therefore remain relatively small and simple.

Finite state machines like HSM's are useful to an extent, but cannot deal with more subtle properties of dialogue systems. One needs more complex dialogue states, and more complex transition rules, than can be encompassed by finite state method. Specification formalisms like StateCharts therefore allow transitions to refer to a more complex state by means of guarded transitions and by allowing transitions to modify this state. Informally, a guard is a boolean valued condition associated with a transition. The transition will become enabled only if the guard evaluates to "true". For our Sera dialogue systems we use this approach in the form of a global information state, and we use so called template based guarded transitions. Here we just make the remark here that the diagrams shown, although not showing the connection with the information state, have been translated into counterparts that do use such guards and state updates. As an example, consider the "u_longSilence" labeled transition within the "userParticipation" diagram. Informally it signifies that the user "has been silent for a long time". Our actual information state machine keeps track of how many seconds our user has been silent, and the transition is "guarded" by a boolean condition that states that this time must be larger than a predefined threshold.

For ease of specification we implemented the HSMTTool, a graphical environment for editing and checking HSMs. In the editor mode new graphs can be made from scratch; models can

be stored in XML format, and reloaded, and adjusted. In the run mode the designer can test his model.

Figure 4 shows the TurnTaking Model as it is used in the SERA system. It extends and refines in a sense the SALTE model of the dialogue system presented in (Wallis 2010). Keep in mind though that its working can only be understood by considering it as part of the whole set of models with which it runs in parallel.

The TurnTaking model:

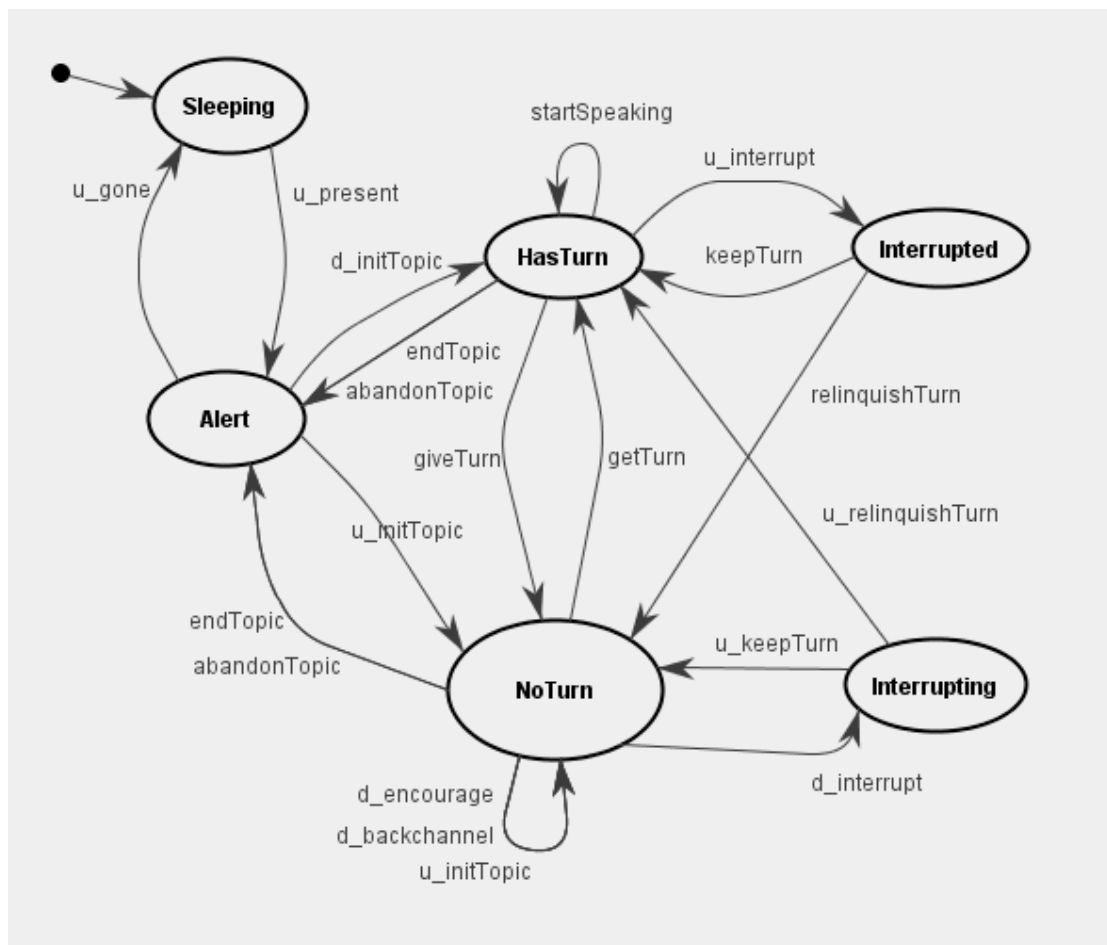


Figure 4 Turn Taking

One of these other models is shown in Figure 5. It is build to cope with the situation in which the user is not attending any more to the conversation.

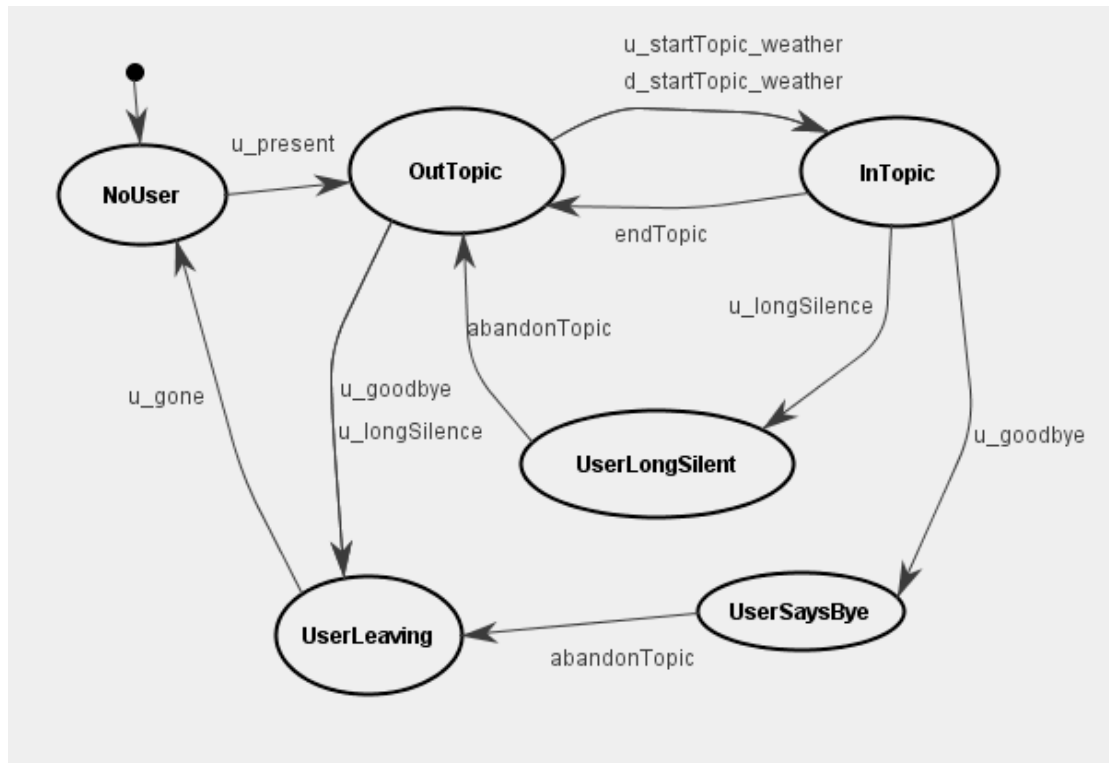


Figure 5 Reacting To User Leaving

When the user is silent for a longer period, the agent will respond with an endTopic. When the user says goodbye, the agent will respond by ending the topic and make the user leave.

4.2 Where does this get us?

In the above, we presented methods and tools for the modular specification, and implementation of topic- and turn-management in dialog systems. We have built a prototype system that can simulate dialogues using these models that allow flexible handling of interruptions and topic switching which are skills that are needed for a system in order to be perceived as social agent. Although we made a step towards building tools for the development of social conversational agents in this way there are some challenges ahead and fundamental questions that remain.

One question that recurs in building intelligent interactive systems is on what level the system should react to changes in the outside world and to feedbacks given by the user: on a cognitive level, implemented by the central Dialog Manager or on a more automatic level, implemented by device specific software or hardware. We have to think further about the criteria for deciding on what level for example feedback or backchannels are processed. In the dialog between P4 and the companion we saw the following fragment.

N4: If you wanted to add some exercise to your log,
 P1: Yep, one hour with my dog this morning.
 N5: then please can I just first quickly tell you the weather and pass on how much exercise you've got planned today.

Our prototype system makes it possible to have an interaction that runs like:

N4: If you wanted to add some exercise to your log,
P1: No, I want to hear about the weather please.
N5: Oh, one moment I give you the weather forecast then.

This could be built using an extended specification that allows the late adaptation after the user's interrupt P1.

A more fundamental question is whether this solves the problems that participant P4 has in aligning her comments with the companion? This requires at least two problems to be solved. Is there a computable set of positions in turns that are potential points of interruption that human users adhere to. That would define the basic units that a speaker turn consists of. Second, is there a fixed set of (multi-modal) signals that a system can use as cues for identifying contributions as interruptions that should be handled in a socially believable way. As long as these problems haven't been solved we believe that P4 should adapt herself to the turn-taking regime of the companion instead of following her own natural turn-taking rules.

5 Conclusion

We have tried to show that in order to build artificial social companions, we need a thorough understanding of the interaction between the human and the machine. In thinking about requirements for social agents one often assumes that one should build systems that act as much human-like as possible to ensure natural interactions. The example dialogue we provided above that people that confronted with machines with limited natural interaction skills have a hard time in deciding whether to treat the machine as a human or to try human-machine interaction strategies that they know from interacting with other machines.

An important goal for the design of more "natural" interaction styles is an understanding of what stance the human takes towards the technology and the other is what the technology assumes about the human in interaction. As designers of interactive systems, we try to predict the actions that the human users will produce in response to the actions of our artefact and we program the artefact accordingly. We hope the human user will build up a mental model that corresponds to the way we programmed the machine. On the other hand, we would like to know how the human users really interact with the machine and what position they assume. To arrive at an understanding of how artificial companions work it is therefore necessary to study real interactions between humans and machines in ecologically valid situations.

Acknowledgements

This work has been supported by the European Community's Seventh Framework programme under agreement no. 231868 (SERA).

6 References

T. W. Bickmore, Relational Agents: Effecting Change through Human-Computer Relationships, Ph.D. Thesis, Massachusetts Institute of Technology, 2003.

O. A. Blanson Henkemans, P. J. M. van der Boog, J. Lindenberg, C. A. P. G. van der Mast, M. A. Neerincx, and B. J. H. M. Zwetsloot-Schonk, An online lifestyle diary with a persuasive computer assistant providing feedback on self-management, *Technology and Health Care*, 17(3):253-267, 2009.

C. Breazeal, C. D. Kidd, A. L. Thomaz, G. Hoffman, and M. Berlin, Effects of nonverbal communication on efficiency and robustness in human-robot teamwork, in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) 2005*, Edmonton, Alberta, Canada, August 2005. IEEE Press, Washington, D.C., pp. 708-713, 2005.

A. Bruce, I. Nourbakhsh, and R. Simmons, The role of expressiveness and attention in human-robot interaction, in *Proc. IEEE International Conference on Robotics and Automation (ICRA '02)*, vol.4, IEEE Press, Washington, D.C., pp. 4138-4142, 2002.

D. Byrne, G. Clore, and G. Smeaton, The attraction hypothesis: do similar attitudes effect anything? *Journal of Personality and Social Psychology* 51:1167–1170, 1986.

J. Cassell, T. W. Bickmore, M. Billingham, L. Campbell, K. Chang, H. H. Vilhjálmsson, and H. Yan, Embodiment in Conversational Interfaces: Rea, in *Proc. CHI, 1999*, ACM Press, New York, NY, pp. 520-527, 1999.

G. Castellano, R. Aylett, K. Dautenhahn, A. Paiva, P. W. McOwan, and S. Ho, Long-term affect sensitive and socially interactive companions, in *Proceedings of the Fourth International Workshop on Human-Computer Conversation*, 2008.

K. Dautenhahn, Design spaces and niche spaces of believable social robots, in *IEEE Int. Workshop on Robot and Human Interactive Communication*. Berlin, Germany, IEEE Press, Washington, D.C., pp.192-197, 2002.

B. R. Duffy, Anthropomorphism and the social robot, *Robotics and Autonomous Systems*, 42(3-4):177-190, 2003.

B. J. Fogg, and C. Nass, Silicon sycophants: the effects of computers that flatter, *International Journal of Human-Computer Studies*, 46(5):551-561, 1997.

T. Fong, I. Nourbakhsh, and K. Dautenhahn, A survey of socially interactive robots, *Robotics and Autonomous Systems*, 42(3-4):143-166, 2003.

J. Goetz, S. Kiesler, and A. Powers, Matching robot appearance and behavior to tasks to improve human-robot cooperation, in *Proceedings of the 12th IEEE International Workshop on Robot and Human Interactive Communication (RO-MAN 2003)*, IEEE Press, Washington, D.C., pp. 55-60, 2003.

N. Green, W. Lawton, and B. Davis, An assistive conversation skills training system for caregivers of persons with Alzheimer's Disease, in T. Bickmore (ed.) *Dialogue Systems for Health Communication, Papers from the 2004 AAAI Fall Symposium*, Technical Report FS-04-04, AAAI Press, Menlo Park, CA, pp. 36-43, 2004.

F. Hegel, M. Lohse, and B. Wrede, "Effects of visual appearance on the attribution of applications in social robotics," in *Proc. of the 18th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2009)*, IEEE Press, Washington, D.C., pp. 64-71, 2009.

D. Johnson, J. Gardner, and J. Wiles, Experience as a moderator of the media equation: the impact of flattery and praise. *International Journal of Human-Computer Studies* 61(3):237-258, 2004.

D. Johnson, and J. Gardner, The media equation and team formation: Further evidence for experience as a moderator. *International Journal of Human-Computer Studies*, 65(2):111-124, 2007.

C. D. Kidd, Designing for Long-Term Human-Robot Interaction and Application to Weight Loss, PhD Thesis, Massachusetts Institute of Technology, School of Architecture and Planning, Program in Media Art and Sciences, 2008.

K. Kobayashi, K. Funakoshi, S. Yamada, M. Nakano, Y. Kitamura, and H. Tsujino, Smoothing human-robot speech interaction with blinking-light expressions, in *Proceedings of the 17th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2008)*, IEEE Press, Washington, D.C., pp. 47-52, 2008.

H. Kozima, C. Nakagawa, and H. Yano, Attention coupling as a prerequisite for social interaction, in *Proc. of the 12th IEEE International Workshop on Robot and Human Interactive Communication (RO-MAN 2003)*, IEEE Press, Washington, D.C., pp. 109-114, 2003.

S. Li, B. Wrede, and G. Sagerer, A dialog system for comparative user studies on robot verbal behavior, in *Proc. of the 15th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2006)*, IEEE Press, Washington, D.C., pp. 129-134, 2006.

R. Looije, F. Cnossen, and M. A. Neerincx, Incorporating guidelines for health assistance into a socially intelligent robot, in *Proc. of the 15th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2006)*, IEEE Press, Washington, D.C., pp. 515-520, 2006.

R. Looije, M. A. Neerincx, and F. Cnossen, Persuasive robotic assistant for health self-management of older adults: Design and evaluation of social behaviors, *International Journal of Human-Computer Studies*, vol. in Press, Corrected Proof, 2009.

H. Prendinger and M. Ishizuka, *Life-Like Characters: Tools, Affective Functions, and Applications (Cognitive Technologies)*. SpringerVerlag, 2004.

B. Reeves and C. Nass, *The Media Equation: how people treat computers, televisions and new media like real people and places*. Cambridge, Cambridge University Press, 1996.

P. J. Sengers, Designing Comprehensible Agents, in *Proceedings of the Sixteenth International Conference on Artificial Intelligence (IJCAI-99)*, Morgan Kaufmann, Los Altos/Palo Alto/San Francisco, pp.1227-1232, 1999.

C. L. Sidner, C. D. Kidd, C. Lee, and N. Lesh, Where to look: a study of human-robot engagement, in *Proceedings of the 9th international conference on Intelligent user interfaces (IUI'04)*, ACM Press, New York, NY, USA, pp. 78-84, 2004.

C. Smith, N. Crook, J. Boye, D. Charlton, S. Dobnik, D. Pizzi, M. Cavazza, S. Pulman, E. Santos de la Camara and Markku Turunen (2010) Interaction Strategies for an Affective Conversational Agent. Proc. 10th International Conference on Intelligent Virtual Agents (IVA2010), 20-22 September 2010, Philadelphia PA, Lecture Notes in Computer Science, 2010, 301-314.

A. Tapus, C. Tapus, and M. J. Mataric, The role of physical embodiment of a therapist robot for individuals with cognitive impairments, in *Proc. of the 18th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2009)*, IEEE Press, Washington, D.C., pp. 103-107, 2009.

P. Wallis, A robot in the kitchen, in *Proc. of the 2010 workshop on companionable dialogue systems*, pp. 25-30, 2010.

M. L. Walters, D. S. Syrdal, K. Dautenhahn, R. te Boekhorst, K. L. Koay, Avoiding the uncanny valley: robot appearance, personality and consistency of behaviour in an attention-seeking home scenario for a robot companion. *Autonomous Robots* 24(2):159-178, 2008.