‘I, Max’ – Communicating with an Artificial Agent

Ipke Wachsmuth

Artificial Intelligence Group, Faculty of Technology, University of Bielefeld
33594 Bielefeld, Germany
ipke@techfak.uni-bielefeld.de

Abstract. With the advent of communicating machines in the form of embodied agents the question gets ever more interesting under which circumstances such systems could be attributed some sort of consciousness and self-identity. We are likely to ascribe to an agent with human appearance and conducting reasonable natural language dialog that it has desires, goals, and intentions. Taking the example of ‘Max’, a humanoid agent embodied in virtual reality, this contribution examines under which circumstances an artificial agent could be said to have intentional states and perceive others as intentional agents. We will link our examination to the question of how such a system could have self-awareness and how this is grounded in its (virtual) physis and its social context. We shall discuss how Max could be equipped with the capacity to differentiate between his own and a partner’s mental states and under which conditions Max could reasonably speak of himself as ‘I’.

Keywords: embodied agents, intentional states, machine consciousness, self-knowledge, emotion, memory

1 Preliminaries

A lot of people talk to their computer – mostly if it doesn’t work as desired. This is certainly by no means true communication with the machine, which need not be explained further. Research into artificial intelligence aims, among other things, at enabling machines (or even machine ‘beings’) to communicate with people as genuinely and naturally as possible. This requires, first of all, machines that are able to perceive and represent their environment, draw conclusions and act accordingly.

Evaluating whether communication between a human being and a machine may actually be possible depends on what is to be understood precisely by ‘communication’. If it is supposed to be transferring information that makes the receiver change its behavior, then even pushing the button that releases the copying machine from stand-by to action may be considered human-machine communication. In this sense the term is, in fact, used in the engineering sciences. If, however, both communicating partners are required to be systems acting autonomously and making use of a common repertoire of signs in order to inform each other or to negotiate deals, this seemed, first of all, rather limited to humans. The communicating partners might even be expected to perceive themselves – as well as each other – as intentional agents, and to be conscious of themselves and the other.

With the advent of communicating machines in the form of embodied agents the question gets ever more interesting whether such systems could have some sort of consciousness and self-identity in a foreseeable future. It is tempting to ascribe an agent which has human appearance and which can conduct reasonable natural language dialog that it has certain beliefs and desires, pursues certain goals, and behaves rationally in the sense that it will act to further its goals in the light of its beliefs. That is, we are likely to conduct such a dialog from the intentional stance [10]. But still, even when our artificial opposite had a name to which it attends and called itself 'I', we would assume that attributing the agent consciousness is inadequate at the given time.

In the case of human beings, the term consciousness describes the fact that we are aware of our thoughts and sensations. Our thinking, feeling and will are – more or less well – available to us and, by way of language, we are even able to communicate this (more or less well) to others. At a closer look, the somewhat colorful concept of consciousness is differentiated into quite different forms[1]. *Firstly*, there is a consciousness of sensations: human beings are aware of the quality of what they experience, e.g. how it feels to touch something or feel pain. *Secondly*, there is a consciousness as being aware of oneself: people know of their physical existence and identity, e.g. they recognize themselves in a mirror. This knowledge is rooted in the perception of one's own body, which we can touch to confirm that we exist and which establishes ourselves in the environment.

And *thirdly*, perception of our physical self, our body and its position in the environment is presumably the basis for our self-perception as an acting being that employs means to pursue goals, even if these are shifted to abstract realms (how to reach my goal = how to get there). This includes being conscious of oneself as a subject of experience, relating one's feelings and thoughts to one's own body and mind and knowing that oneself has caused the effects of actions. This still does not mean, though, that one has to refer to oneself as 'I' or must even have the ability to talk, as will be shown later.

The action perspective, however, is essential to this view, for actions cause changes in the world, whether intentional or unintentional. Actions may have success or fail depending on whether goals aimed at are achieved or not. If an action is successful we are happy, if it isn't we may feel angry. In particular this also applies to communicative actions that constitute the special topic of this contribution. If I tell someone else that 'my knee hurts' this is – different from an involuntary 'ouch' – intentional communicative acting. It is intended to inform the other person about my condition, and I am convinced that he (or she) is able to understand me and my feelings and it is my desire that he should feel sorry for me. I might even expect the other one to offer help.

When communicating with each other, human beings assign each other such 'inner life' (intentional states). Analogous to ourselves, we assume that the other person has intentions, beliefs, desires and goals, which we cannot identify directly, however. We imply that they are there, though, since the other person is a being equally thinking and feeling. And we communicate with the aim of influencing the internal states and thus the actions of the other person. This may be successful or fail. The other person

[1] We shall relate these ideas, put forward here for motivation, to research literature in Section 3.

may stick to her opinion that I am feeling well – although I am telling her that my knee hurts – if she doesn't actually see me limp. Or she believes that my knee hurts although I have only pretended pain, i.e., beliefs may be false.

Human beings possess the ability to recognize in others not only an object of the environment but an acting subject, an image of ourselves, but with its own perspective and intentions. We can even develop a mental model of our communication partner, making assumptions – possibly false ones – about the other's beliefs, desires, and intentions ('Theory of Mind'; see also Krämer [16]). Developing such representation of the other – a 'partner model' – is only possible due to the fact that intentional states have contents that can be expressed in the form of statements (she knows I am very busy, she wants me to come home earlier today, she wants me to go with her to the movies tonight, etc.). Such representation requires symbols of some kind as 'thought signs', which carry contents that form the basis of our logical thinking and rational acting.

Being able to not only understand others as intentional agents but also to reason about their thoughts and goals requires a high degree of consciousness, which – as many researchers believe – is coupled to symbolic representations of the world (but see [2] for discussion). We shall in the following firstly investigate whether and under what circumstances artificial systems can be justifiably attributed intentional states, i.e. can *have* intentional states. Secondly, given certain cognitive conditions, are machines able to *know* about themselves, and are they capable of understanding intentions and perspectives of a dialog partner? Before discussing this, two additional aspects shall be addressed that are closely linked to consciousness, namely, emotion and memory.

As already mentioned above, our feelings play a decisive role when evaluating the success of an action. Even more, emotions are considered to be a basic condition for organized action in modern theories of cognition. Among other things, emotion is understood to be a control medium of the cognitive system to regulate attention directed to incoming stimuli in order to differentiate between important and unimportant matters. The fact that one becomes aware of something is apparently in major parts connected to affective experiencing. Moreover, emotions are of essential importance to the ability of differentiating between various options for actions (see Damasio [9]), as well as for the significance of experiences that affect our memories more permanently. In the case of human beings, storing information is closely connected with the affective appraisal but also, on the other hand, with the realization that the event concerned is very special or rare. This observation indicates that not only emotion but also memory constitutes an important aspect of consciousness.

Our experiencing would be incomplete and the awareness we have of ourselves would not be very profound, if our mind were not equipped to store memories – in particular those that concern ourselves, something we have experienced just before, or experienced yesterday or a long time ago. Generally, we are able to access our personal past, an ability that has been called autobiographical memory; cf. Conway and Pleydell-Pearce [8]. This is the basis of a form of consciousness called 'autonoetic' (knowing of oneself) allowing us to imagine our identity – uncoupled from current experiences – in the past and the future. Studies of patients with impaired consciousness and impaired memory suggest a connection between autonoetic consciousness and memory, in particular episodic memory; cf. Markowitsch [18].

In the section to follow we will introduce ‘Max’ [14], an artificial humanoid agent embodied in virtual reality (see also [13]). To examine the questions raised above, we will then turn to current research discussions of how an artificial system could have consciousness and self-awareness. Finally, we will discuss the conditions under which Max could reasonably speak of himself as ‘I’. As we go along, we shall sketch starting points for the technical realization of such conditions and also discuss the roles of emotion and memory.

2 Who is Max?

“Could you imagine Max being conscious of himself one day?”, I was asked on a conference some time ago. Earlier in the meeting, I had presented our Bielefeld works on a ‘situated artificial communicator’ called Max. Max is an artificial agent (a ‘virtual human’) communicating with his human opposite verbally and through body language, with gestures and mimic. Resembling human appearance he can be met in our laboratory in the setting of a three-dimensional computer graphics projection. Max helps us investigate in detail the basics of communicative intelligence and how it is possible to describe it – in parts – so precisely that a machine can be made to simulate it (the virtual human Max is a program-controlled software-machine). This means that collecting insights about the functioning of human communication is an important focus of our work. A technical goal is also the construction of a system as functional and convincing as possible that may be applied in different ways.[2]

Our research scenario deals with the cooperative construction of objects, such as an airplane from a *Baufix* wooden toykit. This is to test whether Max shows a kind of ‘understanding’ in changing situations that enables him to perform a dialog with a human being. Even if he may not be taken for a human being, Max is intended to show forms of communication that are familiar to us, look into our face when talking or listening, gesticulate naturally, look enquiringly if he does not understand us and – before talking himself – he should wait till we have finished.

The human and Max are opposing each other, for instance at a table (see Figure 1). With the exception of the person shown left, the pictured scene is projected virtual reality. On the table, there are different building bricks, three- and five-hole bars, several screws with colored caps, a cube with holes on all sides etc., all items will be assembled in the course of the dialog. The human sees Max and the entire setting three-dimensionally by means of stereo glasses and hears the synthetic voice of Max due to surround sound from hidden speakers. Max ‘sees’ the human, his gaze direction as well as hand and arm movements by means of an infrared tracking system and cordless data gloves, and – via a radio microphone – he ‘hears’ him speak which he processes with a speech recognizer. Both the human and Max are able to have individual parts assembled by natural language instructions and gestures, carried out in simulation with realistic sounds added. This is done, for example, as follows:

[2] Since January 2004, Max has been performing at the Heinz Nixdorf Museums-Forum in Paderborn where he informs visitors – in a dialog situation – on all things to know about one of the exhibition areas and the computer museum. It is the first time Max came into real-life application.

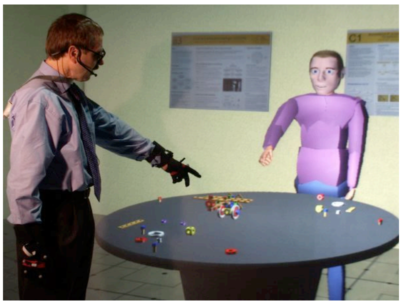


Fig. 1. Cooperative dialog with Max.

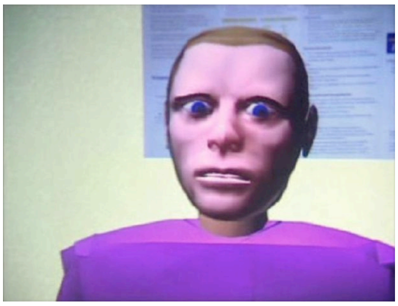


Fig. 2. Max, unpleasantly touched.

Max: Hello, I'm Max. How can I help you? [*smiling*].

Human: Explain the assembly of a propeller.

Max: Let's do it together. First, a screw has to be inserted into the middle of a three-hole-bar.

Human: Insert the yellow screw into the bar [*points to a three-hole-bar, the first screw is inserted into the first hole*].

Max: No, that was the wrong hole [*screw slips out*].

Max: I'll explain it to you again. A screw has to be inserted into the middle of a three-hole-bar.

Human: Insert the screw into the middle of the three-hole-bar [*screw is inserted into the middle of the three-hole-bar chosen before*].

Max: Okay, that's correct. Now insert this into the three-hole-bar [*points to the bar with the screw*].

Human: You do it.

Max: Okay, I will perform the action [*the objects are connected as instructed*].

Max: And now the bars need to be turned crosswise to each other [*shows the rotation using a gesture*] etc.

The cognitive architecture the virtual human Max is based on will be explained in a later section. Just this much: Max knows some grammatical rules and makes use of a semantic lexicon for translating the meaning of words in order to represent the meaning of his opposite's utterances in the form of symbolic descriptions using compositional and reference semantics. Within a limited vocabulary, Max is able to talk and coordinate his gestures due to his flexible body. Simulated facial muscles enable him to express 'emotional conditions' (see Figure 2), which – among other things – are influenced by achieving or failing communicative goals. Max's verbal utterances are produced from a repertoire of stereotype expressions by adapting their parameters to the current situation – including the generation of appropriate gestures. This also includes the term 'I', without Max having a notion of himself (at the current time).

In the theory of communicative action, these dialog expressions could be seen as actions in the proper sense only against the background of attributing intentional states. This means, e.g., Max ought to have a mental state of some kind such as 'wants to have an answer', to make his initial question 'genuine' communication. First of all, Max's body movements, too, are (simulated) physical events. Only in connection with an intended communicative function (as represented in the form of goals) would they gain importance as gestural actions, i.e., only by the fact that a sequence of individual movements is projected and carried out in line with a currently represented mental state of a communicative goal. Seen from the philosophical angle, they would be attributed the status of actions only if Max were able to perform his dialog from the first-person perspective. Would it thus be possible for Max to have that kind of consciousness of his self? Before trying an answer, a brief overview of the state of research into 'machine consciousness' will be given.

3 Consciousness in Artificial Systems?

The question as to whether machines are able to develop forms of consciousness has been a topical subject within artificial intelligence, the neurosciences and, not least, in the philosophy of mind. Research into 'machine consciousness' is expected to yield also further insights on human consciousness. In particular we might find it somewhat strange to attribute a human-like opposite a profound ability to communicate, if he were not able to reasonably speak of himself as 'I'. This would require, however, to configurate the artificial agent accordingly, so as to enable him to adopt a first-person perspective. After outlining a few research approaches towards this subject, different forms of 'self-knowledge' will be discussed in particular.

3.1 Machine Consciousness

Machine consciousness projects can be placed along a spectrum, one of its poles represented by modeling physical brain processes. The digital neuromodels by Igor Aleksander, for instance, are based on the theory that brain cells balance sensory input in a way that allows them to consistently represent real-world objects, in other words, they encode a neuronal depiction of the exterior world; cf. Aleksander, Morton, and Dumall [1]. The other pole is the embedding of preprogrammed rules for controlling the behavior of an artificial agent; e.g., Sloman [26]. Roughly in the middle between both extremes, there is the Global Workspace theory of Baars [3], [4] positing that consciousness emerges if multiple sensory inputs trigger neural mechanisms, which compete to ascertain the most logical response to the inputs. The "Intelligent Distribution Agent (IDA)" model by Franklin and Graesser [12], for instance, is based on this hypothesis.

Research approaches towards modeling mental states and practical reasoning are frequently based on functional models of planning and choosing actions by means-ends analysis, mainly in versions of the *belief-desire-intention* paradigm (BDI); cf. Rao and Georgeff [22]. The BDI approach comes from Michael Bratman [7]; one of its fundamentals can be traced back to the work of Daniel Dennett [10] on the behavior of intentional systems. The basic idea is the description of the internal working state of an agent by means of intentional states (beliefs, desires, intentions) as well as the layout of a control architecture that allows the agent to choose rationally a sequence of actions on the basis of their representations. By recursively elaborating a hierarchical plan structure, specific intentions are generated until, eventually, executable actions are obtained; cf. Wooldridge [29]. Identification and representation of beliefs, desires and intentions are also useful for analyzing the behavior of artificial agents that communicate with humans or other artificial agents; see Rao and Georgeff [23].

Modeling intentional states is based on their symbolic representation. One of its assets is the flexibility it provides for planning and reasoning. In beliefs, for instance, facts concerning the world may be stored that an agent is not (or no longer) able to perceive at the moment, which, however, are to affect his further planning. An agent being able to pursue his goals not only in the light of currently perceived information but also with reference to world knowledge, remembered past and anticipated future will be superior to other agents that do not possess this ability. Even in view of the continuing debate on the significance of symbolic representations for human intelligence it is reasonable to assume that humans represent intentional states symbolically and draw their conclusions on this basis.

It is a difference, though, whether an agent draws conclusions simply on the basis of his beliefs and desires or whether he makes use of them – with a corresponding description – for drawing conclusions, recognizing them to be his own. In many cases such differentiation may not have functional advantages. An agent should be expected, however, to represent his intentional states explicitly as being his own ones, if he must also record and deal specifically with other agents' intentional states. Agents are going to communicate with the intention of changing the inner states of other agents, i.e. their beliefs and intentions. Given a favourable situation, an agent being 'conscious' of his goals may realize them in an opportunistic way.

3.2 Physically Grounded Self-Knowledge (Anderson and Perlis)

From the philosophical point of view, consciousness develops if an agent constructs a model of himself and integrates it into his model of the world [11], [19]. It is a frequently discussed question whether this requires a certain linguistic competence and, in particular, the capacity of using in self-representations an indexical symbol ('I') that refers to oneself. According to Anderson and Perlis [2], usage of an indexical symbol is not imperative for an agent – whether human or artificial – for being able to recognize oneself as the origin of actions. Rather, they argue, it would suffice if the agent had a basal concept of himself rooted in his bodily self-perception which they term *essential prehension* – in opposition to John Perry's [20] well-known problem of the *essential indexical.*

In their initial argument, Anderson and Perlis use the example of the fictitious robotic agent JP-B4 that accumulates information about himself on a self-token[3] 'JP-B4' (he is thus expected to recognize, for instance, that he himself caused an oil stain). This self-token is a self-representation for JP-B4, if especially any physical action performed by JP-B4 which keeps his self-token as a direct object in the

[3] The authors speak of self-referring (mental) token or self-representing (mental) token which is to be understood as a 'marker' of some kind, indicating self-related information.

description of the action, is directed towards himself within the world. To this end they need the assumption that JP-B4 has proprioceptive sensors reporting the spatial position of his limbs and his movable sensors. With this, JP-B4 is able to represent his body as an object (one of many) which, however, is made special due to the fact that the positions of perceived objects (such as the oil stain) can be determined relative to the agent.

In the case of humans, too, Anderson and Perlis go on to argue, perception of one's own body (somatoception) through the tactile sense, proprioception, etc. constitutes the basis of a *physical* self-representation fixed in the environment that is even required for actions as simple as reaching for an object and that is rooted in self-identification. Analogous to JP-B4 they postulate as the only basis for this a special mental self-representing token ('SR*') that is to mark automatically somatoceptive information and that must also be present in mental representations of (initially physical) self-directed actions. This self-token may also serve to relate to oneself externally perceived informations and align those with body perception without the thinking of a self-symbol (indexical thoughts) being required.[4] Finally, Anderson and Perlis argue that intentional and reflexive self-representations are the result of the cognitive system using the same token 'SR*' when representing intentional states,[5] and that a more comprehensive *self-awareness* is rooted herein.

3.3 Implicit and Explicit Self-Knowledge (Beckermann)

Beckermann [6] deals with the problem under what conditions cognitive systems (also artificial ones) – or 'agents', a term that is preferred here – may obtain an explicit form of self-awareness based on reflexive self-knowledge. He supports the thesis that cognitive agents[6] may possess reflexive self-knowledge exactly when they make use of (meta)-representations concerning themselves and that are, in addition, coordinated with 'agent-relative' representations.

Agent-relative knowledge is knowledge represented from the perspective of a particular agent. As long as the agent perceives the world and himself from his own perspective only, he does not need an explicit reference to himself (and, accordingly, no self-symbol) in his representations. Rather, he is able to generate representations on the basis of an implicit reference system in the center of which the agent himself is located. An example: 'The apple nearby that can be reached for' which he is able to grab without thinking 'I'. Neither do sensations such as 'the knee hurts' require the reference to the 'I'. Agent-relative representations thus only include knowledge about the way in which the perceived environment – including the bodily self-perception – is related to the agent. Since they are solely set up from his own perspective, such a representation does not require a self-symbol.

[4] To put it simply: The fact that one sees externally what one feels internally – for instance when touching one's own body – leads to a connection of actions and effects of actions and, thus, to self-identification.

[5] If necessary, they allow the self-token also to be translated as 'I'.

[6] Here, cognitive agents are understood as systems that represent their environment in an internal mental model in order to better cope with their environment.

Now, under what conditions would an agent be forced to introduce an explicit representation of himself? This is discussed by Beckermann as follows, a fictitious agent called 'AL' serving as an example: When representing the perceived environment, AL introduces an internal name for each object – such as 'object-6', 'object-7', etc. – thus representing information about the objects, i.e., their type, properties and relations to other objects. This procedure does not require AL to introduce a name *for himself* – he does not see himself as an object. This becomes inevitable only when AL encounters an object in his environment that he identifies as another cognitive agent. For AL this other agent also constitutes an object for which AL introduces a name – for instance, 'object-111'– whose behavior, however, actually depends on how the other one, for his part, represents the environment.

To be able to predict the behavior of his fellow being, AL has to set up representations of the other's (assumed) representations, i.e. meta-representations – a mental model of the other's mental model. If, for instance, AL believes that the agent he calls 'object-111' considers an item of the environment to be green – e.g., a sofa that AL calls 'object-7' –, or if AL believes that agent 'object-111' desires to sit down on 'object-7' – the green sofa – he sets up agent-relative meta-representations as follows ('believes' and 'wants' in this case concern the intentional states the other agent is assumed to have):

> (believes object-111 (color object-7 green))
> (wants object-111 (sitting-on object-7))

To be capable of representing which (assumed) representations the other one keeps about him, AL is forced to introduce an internal name for himself – such as 'object-100'. Only by means of this name for himself is he able, for instance, to represent adequately the other's desire to obtain food from AL or the other's belief that AL suffers from a hurting knee:

> (wants object-111 (gives-food object-100 object-111))
> (believes object-111 (pain-in-the-knee object-100))

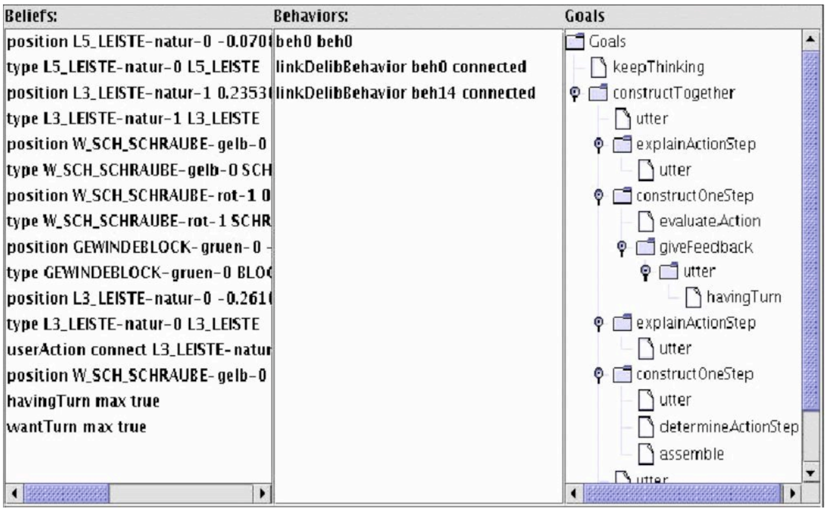
The crucial factor is that AL would now be able to establish a systematic relation between explicit representations that contain this new name (for himself) and his former agent-relative representations with implicit self-reference; i.e. (sitting-on object-7) refers to (sits-on object-100 object-7), meaning: if AL knows he is sitting on the green sofa AL realizes that the agent who is *he himself* is sitting on the sofa. In the same way, AL's body perception could be represented not only in an agent-relative way but also explicitly, i.e. (pain-in-the-knee object-100), etc. And since AL's agent-relative representations correspond solely with his respective 'object-100' representations, the special role of the name 'object-100' as a self-symbol is resulting.

As a further effect, AL would also be able to generate meta-representations *about himself*, thus seeing himself from an external perspective, e.g. (desires object 100 (sitting on object 7)). Only then would he know his own beliefs and desires, could develop explicit self-knowledge and, hence, self-awareness. Only then is it conceivable that AL – together with his fellow beings – develops a language which includes word symbols like 'I' and 'you'. He would have learned the meaning of the word 'I', when he used it to express only such representations that are related to himself, i.e., he says ‘I’ only if he talks about himself.

Hence, explicit self-knowledge (i.e. representations with a name for oneself) develops only in the social context: if a cognitive agent meets other cognitive agents and he realizes that – just as he does – they represent their environment and thus him, too.[7] If the agent desires to represent for himself such representations of his fellow beings of which he is the object, he is forced to introduce his own internal name and make himself explicity an object of his representations. If he finally takes the step to bring his agent-relative representations in alignment with their explicitly self-related counterparts, he has got reflexive self-knowledge.

4 Max as a Cognitive Agent

Let us turn back to Max. Max is no fictitious robot but a fully implemented system that designs a humanoid agent in virtual reality. He is equipped with an articulate flexible body which – among other things – allows him to access parameters of his physis in order to enable him, e.g., to call up his position in the environment and his joint angles when planning his gestures; cf. Kopp and Wachsmuth [15]. As already mentioned above, our scenario deals with dialogs between a human and Max in the course of which a model airplane will be constructed.



Beliefs:

```
position L5_LEISTE-natur-0 -0.070
type L5_LEISTE-natur-0 L5_LEISTE
position L3_LEISTE-natur-1 0.2353
type L3_LEISTE-natur-1 L3_LEISTE
position W_SCH_SCHRAUBE-gelb-0
type W_SCH_SCHRAUBE-gelb-0 SCH
position W_SCH_SCHRAUBE-rot-1 0
type W_SCH_SCHRAUBE-rot-1 SCHR
position GEWINDEBLOCK-gruen-0 -
type GEWINDEBLOCK-gruen-0 BLO
position L3_LEISTE-natur-0 -0.261
type L3_LEISTE-natur-0 L3_LEISTE
userAction connect L3_LEISTE-natur
position W_SCH_SCHRAUBE-gelb-0
havingTurn max true
wantTurn max true
```

Behaviors:

```
beh0 beh0
linkDelibBehavior beh0 connected
linkDelibBehavior beh14 connected
```

Goals

```
Goals
  keepThinking
  constructTogether
    utter
    explainActionStep
      utter
    constructOneStep
      evaluateAction
      giveFeedback
        utter
          havingTurn
    explainActionStep
      utter
    constructOneStep
      utter
      determineActionStep
      assemble
    utter
```

Fig. 3. Present beliefs, behaviors, and goals of the Max system (from Kopp et al. [14]).

[7] A bit trenchant: Hermits would be able to manage with agent-relative representations, i.e. implicit self-reference.

As a cognitive agent, Max represents his (virtual) world in parts to be able to cope with the tasks when assisting the (virtual) construction of Baufix objects. For each Baufix object – either existing from the beginning or implemented later, e.g. aggregated ones – he introduces a formal internal name such as 'object-1', 'object-2', etc. (It is not relevant here that within the system actually a symbol generator provides somewhat more differentiated 'talking names', see Figure 3). Furthermore, he records beliefs about the type of the parts and their position giving the vectors of a reference point of an object within the world coordinate system, for example as follows:

```
(type object-1 THREEHOLEBAR)
(position object-1 (2,3,5))
(type object-2 THREEHOLEBAR)
(position object-2 (x,y,z))
(type object-3 SLOTHEADSCREW-yellow)
(position object-3 (x',y',z'))
```

Changes of the scene are represented by Max real-time, for instance, by asserting (connected object-26 object-27) when the according parts are connected. Intentional states of his dialog partner have so far not been represented by Max. He represents, however, whose turn it is to speak. Various routines enable Max to identify turn signals of his dialog partner (turn-taking) and to know whether it is his turn or whether he wants to have it (having-turn Max true, want-turn Max true; see Figure 3).

To organize the complex interplay of sensory, cognitive and actoric abilities, a cognitive architecture has been developed for Max [17], aiming at making his behavior appear believable, intelligent, and emotional. Here, 'cognitive' refers to the conception of structures and processes underlying mental activities. Fitted to his trial scenario, Max is equipped with limited knowledge of the world and is capable of planning and reasoning so that he may act as an intelligent assistant. Moreover he is equipped with reactive behaviors that enable him cope with disruptions and sudden changes.

In a hybrid system architecture the *Max* system integrates symbol-processing and behavior-based approaches concerning perception, reactive behavior, higher mental processes such as reasoning and planned action, up to and including focused attention and action appraisal. The central part is a *belief-desire-intention (BDI)* interpreter. Due to the hybrid architecture Max is both able to conduct a dialog with planned utterances and to produce spontaneous utterances, e.g. in the form of turn-taking and feedback signals. Additionally, specialized planners – e.g., for constructing Baufix objects – and specialized memory stores – e.g., with dynamically updated representations for the state of constructed objects have been integrated.

Explicitly represented goals (*desires*), which may be introduced through internal processing and through external influences as well, are serving as inner motivation that is triggering behavior. Intentions are generated by means of the BDI interpreter, which determines the current intention on the basis of existing beliefs, current desires and goals as well as the options for actions. Max can have several desires, the highest-rated of which is selected by a utility function to become the current intention. Options for actions are available in the form of abstract plans that are described by preconditions, context conditions, consequences that may be accomplished, and a priority function. If a concrete plan drawn up on the basis of these facts has been executed successfully, the related goal will become defunct.

The conduct of the dialog is based on an explicit modeling of communicative competences that are related to multimodal communicative acts [21] generalizing the speech act theory by Searle and Vanderveken [25]. Communicative acts are modeled as action-plan operators. The dialog is performed in accordance with the mixed initiative-principle, this means, for instance, that in case the human fails to answer, Max himself takes the initiative and acts as the speaker. The plan structure of the BDI module makes it possible to implement new goals during the performance of an intention that may replace the current intention, provided it has a higher priority. If the previous intention is not specifically abandoned in this process and its context conditions are still valid, it will become active again after the interruption.

Max's behavior is further influenced by (simulated) emotions, which determine as system parameters in which way Max performs actions. The emotive system is, on the one hand, fed by external stimuli (Max's virtual physis, e.g., has touch-sensitive areas), and, on the other, by the cognitive system: reaching or failing to reach main goals generates positive or negative emotions, respectively, that affect the valences of mood of the emotional system, which, in turn, control Max's unintentional external behavior. The emotional expression in Max's face and voice caused thereby may convey feedback-signals to his opposite. In parallel action, mood valences occurring continuously in a three-dimensional abstract space are categorized and symbolically represented as explicit beliefs; so they may take effect when choosing between options for actions; cf. Becker et al. [5]. Max is also able to utter verbally symbolically represented emotional states ('I am angry now'); in this sense, Max *seems* to be 'conscious' of them.

Now, what about the consciousness Max might actually have of himself as a subject? Insofar as his cognitive abilities are based on a BDI architecture, Max can be justifiably ascribed mentalistic attributes that may be characterized by terms such as knowledge, belief, intention. We may thus state, as an intermediate result, that Max not only represents his environment, he has also – in the discourse of the cooperative situation (constructing with Baufix) – intentional states (beliefs, desires/goals, and intentions), which constitutes him as being an intentional agent. But can – or could – he *know* about his intentional states and those of his dialog partners (more precisely, those he assumes his dialog partners to have)? A model enabling Max to make use of the corresponding meta-representations has so far been rudimentarily realized. Up to now it is only concerned with following up the role of the speaker as well as turn-taking. This is an example, however, which can already serve to indicate to what extent Max requires reflexive knowledge.

As already mentioned above, Max is able to identify turn signals of his dialog partner, i.e. that the other wants to take the role of the speaker (e.g., if the human interrupts him directly, says 'Max!' or raises his hand). Max actually represents his role as the speaker already with a self-symbol (having-turn Max true), even if this is unnecessary in a dyadic setting; an agent-relative representation would be entirely sufficient: (having-turn true) – or (having-turn false) if it is the other's – the human's – turn. Our next plans include to enable Max to have a reasonable conversation with more than one partner, thus he should be able to keep an account of who of those participating is speaking or wants to do so. It could be expected that for this he uses symbol names for his partners (having-turn Other-1, having-turn Other-2, etc.). But does this actually require him to have a self-symbol, as in (having-turn Max)? Even if this social situation suggests recording the turn-holder by name, Max may still manage without a self-symbol, namely, by representing (having-turn true/false) whether he or someone else is speaking and by differentiating the others by name. To know explicitly 'It's my turn', however, Max ought to have a self-symbol.

Let's look at the (as yet fictitious) situation where three agents – Other-1, Other-2, and Max – are having a conversation and are taking turns. As long as Max 'wants' to have his turn only to say something (want-turn true) he will just have to wait for a suitable opportunity. However, the conversational situation of explicitly passing the speaker-role also may occur (turn-giving), signalling to a direct addressee and to be understood as a call for action, i.e. taking the turn; cf. Sacks et al. [24]. In this case, however, Max should be in a position to recognize that he himself is the addressee and, for instance, represent (wants Other-2 (give-turn Other-2 Max)), etc. In other words, Max would then have – or require – some form of explicit self-awareness that allows him to differentiate between his own and a partner's mental states.

5 Criteria for a 'Human' Non-Human Consciousness

In this paper we examined under what conditions an artifical agent may be able to communicate with some sort of consciousness of being an intentional agent, the agent Max embodied in virtual reality serving as an example. In particular, we asked what kind of cognitive conditions are required to enable Max to know of himself and understand intentions and perspectives of a dialog partner. May – some day – Max justifiably talk about himself as 'I, Max'? But also: being a 'virtual human', might Max constitute a communication partner acceptable for humans?[8]

Let's return to the forms of consciousness first differentiated in the introduction, i.e. (1) consciousness of sensations, of the phenomenal quality of experiencing, (2) consciousness as knowledge of the physical identity, and (3) consciousness in the form of self-perception as an acting being, up to the self-perception as causing actions (here in particular: communicative actions). Let's now consider to what extent that may seem attainable for the artificial agent Max.

1. Qualia. Max can certainly not be ascribed sensations of the kind human beings have, since – due to his virtual body – he does not possess a neurophysiological basis required for qualitative experiencing. In this sense, the simulated emotional states can, for instance, not be experienced subjectively, this means that Max *does not* have feelings. Their functional role in the sense of behavior controlling appraisal, however, can be and has been modeled, at least to some extent – a mechanism of appraisal analogous to feelings, which, for instance, allows a differentiation between options for actions; see Stephan [27]. By means of such appraisal achieved through simulated emotions, Max might be able to develop preferences and directed attention. A positive

[8] An indication that this question is not entirely odd is the fact that Max is frequently asked by visitors of the Heinz Nixdorf Museums-Forum 'Are you a human being?'.

or negative appraisal in achieving or failing communicative goals could to some extent be compared with emotional experiences.

2. Self-identification. Regarding a consciousness as knowledge of the physical identity, the situation is quite different. This concerns the question as to whether Max can have a basal concept of himself that is rooted in the self-perception of his (virtual) physis (*essential prehension* in the sense of Anderson and Perlis [2], see Section 3.1). Let's imagine the following experiment: In virtual reality, Max perceives his simulated – but not yet recognized – mirror image that moves exactly as Max does, e.g., when he places his hand on his left cheek (in our experimental system, touching Max's left cheek gives Max's emotive system a pleasant 'feeling'). It appears technically possible that Max can align the action observed externally and perceived internally by means of a self-token as outlined in Section 3.2. This *physically grounded* self-token mediates an essential awareness of location that is important for acting in space, it is reference point for agent-relative representations and may be the starting point for references a self-symbol makes to the own 'person'.

3. Self-perception as an acting being, even including self-perception of having caused actions (here, in particular, communicative actions): For this, it is mandatory that Max has symbolic representations of his environment and knowledge about how to represent planned actions as goals. Max should in particular be able to represent those as being his goals for which, according to Anderson and Perlis [2], a self-token would suffice. This makes only sense altogether, however, if it can be interlinked. Only through self-identification of his own physis would Max be able to relate agent-relative knowledge *to himself*, by coupling it to a self-token (which is hence grounded in his physis). Only then could he perceive himself as the origin of actions. Only by introducing a self-token into the representations of actions could he perceive himself as causing actions, i.e. establish causal relations between his actions and the effects they trigger.

On the basis of the ideas explained in Section 4, Max *can* be understood as a system perceiving and representing his environment and drawing conclusions in order to cope with the situation. It seems, in fact, relatively easy to construct the *Max* system in a way that all his agent-relative representations are automatically marked *as his own* by means of a self-token. According to the above thoughts, however, conscious self-awareness is coupled to *explicit* self-knowledge, that is, Max would require explicit self-representations incorporating a symbol name for himself, which express Max's view onto himself from an external perspective. To reach this, first of all representational states of the agent on their part must be made the object of representations, i.e. meta-representations need to be set up:

4. Meta-representations. Clearly, by means of the conditions created by a BDI architecture (see Section 4) it appears possible to configurate Max in a way that enables him to set up meta-representations. A more difficult question is how to create an experimental situation that allows Max to set up a *reflexive* meta-representation. According to Beckermann (see Section 3.3), a social situation may be a suitable basis on which Max sets up assumed representations of a communication partner that deal with himself and that he would have to coordinate with corresponding Max-relative representations just as described. As a first step, a turn-taking situation as explained above appears suitable. In this context, Max must be able to set up also expressions in symbolic representations that allow him to make propositions on propositions, even including propositions on himself, such as 'the other wants me, Max, to take the turn'. Max should then be able to derive an according intention that makes him take the turn.

Let's suppose we succeeded in fulfilling such conditions (at least 2-4) for Max altogether. Then we would have to admit Max to be able to communicate as an intentional agent that has self-perception as an acting being. Just as Max would have representational states bearing his intentions, desires and goals, he would be able to ascribe such states to humans. And vice versa, it would then be entirely justified if a human being ascribed him, too, intentions, desires and goals that Max relates to himself, i.e., that are his own.

Yet, the fact that Max's knowledge of his own states would be limited to the moment only would remain rather unsatisfactory; this could still not be considered profound knowledge of himself (autonoetic consciousness). If Max did not know what he did yesterday or what he could be doing tomorrow, he would not have an 'I' persisting across time. A further important criterion thus is:

5. Memory. Max should not only be able to remember who (or whether he) triggered an action, he should also be able to recognize if an event is absolutely new to him, that he cannot remember having experienced before. In order to be aware of being confronted with a new event for the first time, Max should be able to have access to his personal history. He would need an autobiographical memory that enables him, – with reference to his communication partners – to ascertain, 'Yesterday I constructed a plane with you for the first time' or 'I have often (or never) constructed a plane with you'. This requires him, however, to be able to store memories of such an event in appropriate form. If it happened to him a repeated time, it should be possible to revise the uniqueness of the memory, up to daily experience.

How to realize such an autobiographical memory in the case of Max? As described above (Section 4) the behavior-triggering impetus of Max is based on explicitly represented goals. As a starting point for an autobiographical memory, Max may set up a record in appropriate form (marked with a time stamp and his self-token), when one of his goals has been achieved, or failed. It would probably not be helpful if Max stored a record for any processed (sub)goal; there would be far too many marginal ones that are of only temporary importance. Rather, the goals should be evaluated with respect to their significance. Such can be done by the emotive system, by coupling any goal reached and any goal failed with a positive or negative emotion (pleasure or anger), with 'higher' goals triggering stronger and sub-goals less emotional reactions. Permanency of storing memories can be made dependent on the strength of the emotional reaction, thus assuring that memories of main goals remain more pronounced. An adjustment of new and recorded former goals could, in turn, be evaluated emotionally. A goal that frequently failed and has now been achieved for the first time could give rise to Max's 'joyful excitement' and lead to a lasting, 'I'-related memory.

We now see the following picture of artificial consciousness developing: criteria required are self-identification, self-perception as an acting being, meta-representations and memories related to emotional appraisal. Given these conditions, it appears possible that Max approaches forms of 'human' (comparable to a human being's) consciousness and self-identity. The higher the degree of similarity, the more justifiably Max could talk of himself as 'I, Max', and the better Max might be acceptable as a social partner, as a 'human machine' for human beings.

Acknowledgments. This paper is a slightly polished version of an article originally published in German [28]. Assistance by Marion Kämper in providing the English translation and helpful comments by two reviewers and by Manuela Lenzen are gratefully acknowledged.

References

1. Aleksander, I., Morton, H., Dunmall, B: Seeing is Believing: Depictive Neuromodelling of Visual Awareness. In: J. Mira, A. Prieto (eds.): Connectionist Models of Neurons, Learning Processes and Artificial Intelligence. Lecture Notes in Computer Science, Vol. 2084. Springer-Verlag, Berlin Heidelberg New York (2001) 765–771
2. Anderson, M.L., Perlis, D.: The Roots of Self-Awareness. Phenomenology and the Cognitive Sciences 4 (2005) 297–333
3. Baars, B.J.: In the Theater of Consciousness. Oxford University Press, Oxford (1997)
4. Baars, B.J., Franklin, S.: How Conscious Experience and Working Memory Interact. Trends in Cognitive Science 7 (2003) 166–172
5. Becker, C., Kopp, S., Wachsmuth, I.: Simulating the Emotion Dynamics of a Multimodal Conversational Agent. In: E. André, L. Dybkjaer, W. Minker, P. Heisterkamp (eds.): Affective Dialogue Systems. Lecture Notes in Computer Science, Vol. 3068. Springer-Verlag, Berlin Heidelberg New York (2004) 154–165
6. Beckermann, A.: Self-Consciousness in Cognitive Systems. In: Ch. Kanzian, J. Quitterer, E. Runggaldier (eds.): Persons. An Interdisciplinary Approach. ÖBV & HPT, Wien (2003) 174–188
7. Bratman, M.: Intention, Plans, and Practical Reason. Havard University Press, Harvard (1987)
8. Conway, M.A., Pleydell-Pearce, C.W.: The Construction of Autobiographical Memories in the Self-Memory System. Psychological Review 107 (2000), 261–288
9. Damasio, A.R.: Descartes' Error. Emotion, Reason, and the Human Brain. Putnam, New York (1994)
10. Dennett, D.C.: The Intentional Stance. MIT Press, Cambridge, MA (1987)
11. Dennett, D.C.: Consciousness Explained. Little Brown, London (1991)
12. Franklin, S., Graesser, A.: A Software Agent Model of Consciousness. Consciousness and Cognition 8 (1999) 285–305
13. Kopp, S., Allwood, J., Grammer, K., Ahlsén, E., Stocksmeier, T.: Modeling Embodied Feedback with Virtual Humans; same volume
14. Kopp, S., Jung, B., Leßmann, N., Wachsmuth, I.: Max – A Multimodal Assistant in Virtual Reality Construction. KI – Künstliche Intelligenz 4/03 (2003) 11–17
15. Kopp, S., Wachsmuth, I.: Synthesizing Multimodal Utterances for Conversational Agents. Journal Computer Animation and Virtual Worlds 15 (2004) 39–52
16. Krämer, N.C.: Theory of Mind as a Theoretical Prerequisite to Model Communication with Virtual Humans; same volume
17. Leßmann, N., Kopp, S., Wachsmuth, I.: Situated Interaction with a Virtual Human – Perception, Action, and Cognition. In G. Rickheit, I. Wachsmuth (eds.): Situated Communication. Mouton de Gruyter, Berlin (2006) 287–323
18. Markowitsch, H.J.: Autonoetic Consciousness. In: T. Kircher, A. David (eds.): The Self in Neuroscience and Psychiatry. Cambridge University Press, Cambridge (2003) 180–196

19. Metzinger, T.: The Subjectivity of Subjective Experience: A Representationalist Analysis of the First-Person Perspective. In T. Metzinger (ed.): Neural Correlates of Consciousness – Empirical and Conceptual Questions. MIT Press, Cambridge, MA (2000) 285-306
20. Perry, J.: The Problem of the Essential Indexical. In: J. Perry, The Problem of the Essential Indexical and Other Essays. Oxford University Press, Oxford (1993) 33–52
21. Poggi, I., Pelachaud, C.: Performative Facial Expression in Animated Faces. In: J. Cassell, J. Sullivan, S. Prevost, E. Churchill (eds.): Embodied Conversational Agents. The MIT Press, Cambridge MA (2000) 155–188
22. Rao, A.S., Georgeff, M.P.: Modeling Rational Agents within a BDI-Architecture. In: J. Allen, R. Fikes, E. Sandewall (eds.): Principles of Knowledge Representation and Reasoning. Morgan Kaufmann, San Mateo CA (1991) 473–484
23. Rao, A.S., Georgeff, M.P.: BDI agents: From Theory to Practice. Proceedings of the First International Congress on Multi-Agent Systems (ICMAS-95). San Francisco (1995) 312–319
24. Sacks, H., Schegloff, E.A., Jefferson, G.: A Simplest Systematics for the Organization of Turn-taking for Conversation. Language 50 (1974) 696–735
25. Searle, J.R., Vanderveken, D.: Foundations of Illocutionary Logic. Cambridge University Press, Cambridge (1985)
26. Sloman, A.: What Sort of Control System is Able to have a Personality? In: Trappl, R. and Petta, P. (eds.): Creating Personalities for Synthetic Actors. Springer-Verlag, Berlin Heidelberg New York (1997) 166–208
27. Stephan, A.: Zur Natur künstlicher Gefühle. In: A. Stephan, H. Walter (eds.): Natur und Theorie der Emotion. Mentis, Paderborn (2003) 309–324
28. Wachsmuth, I.: “Ich, Max” – Kommunikation mit künstlicher Intelligenz. In: Ch.S. Herrmann, M. Pauen, J.W. Rieger, S. Schicktanz (eds.): Bewusstsein: Philosophie, Neurowissenschaften, Ethik. Wilhelm Fink Verlag (UTB), München (2005) 329–354
29. Wooldridge, M.: Intelligent Agents. In: Gerhard Weiss (ed.): Multiagent Systems – A Modern Approach to Distributed Atrificial Intelligence. The MIT Press, Cambridge MA (1999) 27–77