A lesson from subjective computing: autonomous self-referentiality and social interaction as conditions for subjectivity

Patrick Grüneberg[1] and **Kenji Suzuki**[2]

Abstract. In this paper, we model a relational notion of subjectivity by means of two experiments in subjective computing. The goal is to determine to what extent a cognitive and social robot can be regarded to act subjectively. The system was implemented as a reinforcement learning agent with a coaching function. To analyze the robotic agent we used the method of levels of abstraction in order to analyze the agent at four levels of abstraction. At one level the agent is described in mentalistic or subjective language respectively. By mapping this mentalistic to an algorithmic, functional, and relational level, we can show to what extent the agent behaves subjectively as we make use of a relational concept of subjectivity that draws upon the relations that hold between the agent and its environment. According to a relational notion of subjectivity, an agent is supposed to be subjective if it exhibits autonomous relations to itself and others, i.e. the agent is not fully determined by a given input but is able to operate on its input and decide what to do with it. This theoretical notion is confirmed by the technical implementation of self-referentiality and social interaction in that the agent shows improved behavior compared to agents without the ability of subjective computing. On the one hand, a relational concept of subjectivity is confirmed, whereas on the other hand, the technical framework of subjective computing is being theoretically founded.

1 INTRODUCTION

The mental phenomenon called ‘subjectivity’ has been up to present days one of the central topics of philosophical discussion. Even before the proclamation of the ‘subject’ as a principal of knowledge, one might regard the relation of an epistemic and practical agent to the world and itself as one of the most notorious issues even in antique and medieval philosophy. However, in these days, ‘subjectivity’ enjoys great popularity as phenomenal consciousness, as the individual first-person perspective.[3] But ‘subjectivity’ needs not necessarily be related to consciousness. Instead, recent developments in robotics show that ‘subjectivity’ can also be related to intelligence. Actually, the idea to analyze ‘subjectivity’ as intelligence is not that new [8], [9]. One obvious advantage of decoupling ‘subjectivity’ from consciousness is that intelligence can be analyzed without making use of the most difficult concepts of (phenomenal) consciousness. From this perspective, ‘subjectivity’ is conceptualized as a relational concept, i.e. subjectivity comprises certain relations of an agent towards itself and its environment [16]. The question of phenomenal consciousness is then subordinated in favor of the agent’s self-relation and relations to others. An agent then is supposed to be subjective if it exhibits autonomous relations to itself and others, i.e. the agent is not fully determined by a given input but is able to operate on its input and decide what to do with it. This relational perspective also allows us to take into account social relations. Accordingly, intelligence is not solely a product of internal processes but is constituted in the course of social interaction and therefore builds on the notions of *Aufforderung* [2] and recognition [18].

To narrow down and as an attempt to verify this philosophical and quite abstract notion of subjectivity, we refer to two experiments in subjective computing [17], [14]. Subjective computing aims at utilizing insights from human subjectivity in general, the perceptual process, and human-human interaction for the design of algorithms and human-robot interaction (this concept was initially proposed in [21]). From an engineering perspective it is not the goal to give an account on subjectivity, but rather to utilize certain aspects of human cognition to solve specific problems. One major problem concerns reinforcement learning (RL). Even if an agent is able to decide autonomously about the modification of its behavior, the agent still has to learn what kind of behavior is well suited in order to accomplish a certain task. In order to evaluate the agent’s behavior a coaching function has been implemented into a RL agent so that the agent can receive a trainer’s feedback. The crucial point regarding the potential subjectivity of this agent is that this feedback does not modify the agent’s behavior directly, but the agent interprets the feedback and decides about the subsequent modification of its behavior by itself. Thus, with regard to the relational notion of subjectivity, the agent relates to itself while interpreting feedback and at the same time socially relates to a human trainer. For the implementation of this robotic system both relations, the self-relation in the course of interpretation and the relation to others (the human trainer) enable the robotic agent to successfully accomplish a difficult learning task.

In our relational analysis of the robotic agent we draw upon the ascription of mental abilities to the robotic agent that is supposed to observe, interpret, and reflect the feedback. The question is to what extent this mental behavior is finally algorithmically implemented. By means of an analysis of levels of abstraction [10], we relate the algorithmic to the mentalistic level. The focus lies on an intermediate relational level where it can be shown that the robotic agent exhibits an autonomous self-relation and a social relation to others.

[1] Artficial Intelligence Laboratory, University of Tsukuba, Japan; Institut für Philosophie, Literatur-, Wissenschafts- und Technikgeschichte, Technical University Berlin, Germany, e-mail: patrick.grueneberg@tu-berlin.de.

[2] Center for Cybernics Research, University of Tsukuba, Japan, Japan Science and Technology Agency, Japan, e-mail: kenji@ieee.org.

[3] [24] and [19] are just two of the most prominent examples for phenomenally based accounts. [22] follows a cognitive science approach to subjectivity in terms of a phenomenal perspective to the world and the agent itself.

Even if the robotic agent cannot be conceived of as a full-blown subject (compared to humans), the successful implementation of autonomous self-referentiality and a social relation to others allows us to ascribe subjective states to the robotic agent. Even if the robotic agent cannot be regarded as a full-blown human subject, the successful engineering implementation of this relational structure can be seen as a confirmation for the philosophical notion of subjectivity as increased intelligent behavior has been gained.

In section 2 we will start by shortly introducing the relational concept of subjectivity and by explaining our case study, "coaching a robot based on human affective feedback". After describing the algorithm and the coaching function as well as the experimental layout and results, in section 3, we introduce Floridi's method of levels of abstraction by explaining what this method consists in and how we use it to analyze the robotic agent. By means of four levels of abstractions we will then analyze the robotic agent with a focus on its relational structure. Section 4 begins with an informal analysis that is followed by a formal treatment of the levels of abstraction and their relations. Finally, we evaluate to what extent mental subjective abilities can be ascribed to the robotic agent.

2 PHILOSOPHICAL AND TECHNICAL BACKGROUNDS

2.1 Relational subjectivity

To introduce the concept of relational subjectivity, it is helpful to refer to the current philosophical debate, especially to phenomenal and first-person accounts of subjectivity which enjoy great popularity (see note 3). According to the general idea of this framework, subjectivity consists in phenomenal consciousness so that a phenomenal subject is able to experience its mental states. These mental states refer to environmental entities (or their features, respectively) or states of the agent itself. The experience of these states is subjective to the extent that this experience is only accessible for the agent who has it. Furthermore, such accounts are often representationally or, at least, realistically based, i.e. the experience refers to an objectively existing (independently of the agent) world that becomes conscious in the agent's mind. Although there are plenty of different versions of phenomenal and first-person, representational and realistic accounts, the crucial point for our investigation lies in decoupling subjectivity from the any kind of phenomenal consciousness (for further explanation of the methodological arguments for this decoupling see [16]).

Instead, subjectivity can be grounded in action. In a Kantian and therefore transcendental perspective, this action is conceived of as a condition of the possibility of subjectivity. The main purpose of this action is to structure and construct the subject's reality by means of schematic capacities.[4]. These schematic capacities generate the subject's attitude towards a given reality in which the subject can act. Hence, subjectivity has, secondly, to be decoupled from the notion of a *psychological* subject. The distinction between different individual subjects is not based on different individuals. Instead, the schematic processes make the difference as these exhibit necessary features that apply for every individual subject. Accordingly, subjects usually share the same space-time dimension. On the other hand, schematic processes are not completely determined and thus allow for voluntary action that depends on individual decisions, e.g. on the individual use of cognitive abilities as perception or action. An individual subject can voluntarily focus its visual attention to a certain position in space and decide to move in this direction or to hold its current place. In turn, these voluntary actions depend on determinations that are out of reach for the individual subject, i.e. when visual attention has been focused to a certain position, then the content of the visual experience is determined.

Accordingly, subjectivity is relationally generated by simultaneous processes of determining and voluntary schematic activity. One and the same cognitive action underlies this twofold schematism so that subjectivity is conceived as a relational momentum that is generated in opposition to an objective or determining momentum in an agent's information space. This twofold structure also applies for the individual agent that acts in the social context of other individual agents: On the one hand, the agent relies on its autonomous capacities. At the same time, it depends on social interaction as social interaction constrains its autonomy and therefore provokes a reaction. A reaction here is understood as a self-determination of the agent's actions provoked by some external constraint. Again, a subjective agent is conceived of as relationally constituted. This mutual interdependency of voluntarily determining and necessarily being determined forms the basic framework for a relational concept of subjectivity. In the following we are going to investigate two experiments in cognitive and social robotics in order to evaluate if and to what extent this relational concept of subjectivity can be computationally modeled and implemented. This serves to narrow down and concretize the quite abstract relational notion; at the same time, the framework of subjective computing can be made more explicit; especially we hope to clarify what it can mean for a robotic agent to behave subjectively.

2.2 Case study: coaching a robot based on human affective feedback

Generally, interaction and henceforth social intelligence are regarded as a constitutive part of intelligence at all [5]. Based on an interactive learning algorithm reciprocal interaction between a robotic agent and a human instructor is facilitated. This way of situated learning enables the coach to scaffolding acts of providing feedback [23], while the robot demonstrates its mastery of the task continuously by means of improved behavior. In this kind of peer-to-peer human-robot interaction the robotic agent has to perceive emotions and learn models of its human counterpart [11]. Hence the robot needs to be at least socially receptive, i.e. socially passive in order to benefit from interaction, or coaching respectively [1], and socially embedded, i.e. situated in a social environment and interacting with humans. If the agent is structurally coupled with the social environment, he will be able to be partially aware of human interactional structures [7]. In order to socialize robots have to be compatible with human's ways of interacting and communicating. On the other hand humans must be able to rely on the robot's actions and be allowed to have realistic expectations about its behavior.

In the context of embodied cognition, we are able to model subjectivity as an interactional (social) and therefore relational issue. This means that subjectivity is realized in the course of social interaction which is investigated in the field of social robotics. One core issue in designing social robots consists in socially situated learning. New skills or knowledge are acquired by interacting with other agents. Beside robot-robot interaction (so-called "swarm intelligence" or "collective intelligence"), human-robot interaction displays another major approach [6], [12]. We focus on the case of teaching a robot [28], [29] by means of coaching. Unlike teaching the coaching process does not depend on an "omniscient" teacher that guides the agent toward the goal, but the instructor only gives hints and clues in terms

[4] See historically [20] and, in the sense of an extended schematism, [9]; in the following we refer to an updated schematic account in [16]

of a binary feedback, i.e. positive or negative. It is then the robot's cognitive task to process this feedback and control its actions autonomously.

Our approach to subjective computing is based on two experiments on coaching a robot. These experiments were conducted at the Artificial Intelligence Laboratory (University of Tsukuba) previously to this investigation. The coaching process itself bears on two relational aspects that are the focus in these experiments:

1. the cognitive process of autonomous interpretation of the feedback by the agent [17]
2. the social interaction between the human instructor and the robot [14]

In the following we will, firstly, describe the problem that underlies the implementation of the coaching RL agent and of affective feedback, respectively. Secondly, we illustrate the experimental setups and results.

The first experiment [17] was conducted by Hirokawa and Suzuki and consists in a reinforcement learning (RL) agent with an implemented coaching function so that the robotic agent is open to human feedback during its behavior learning. While coaching had already been implemented before [25], [26], RL offers a significant advantage. A coaching RL agent is able to learn automatcially by its own internal values. RL is a commonly used method for autonomous machine learning based on the idea that an agent autonomously adapts to the specific constraints of an environment [27]. While often a learning algorithm is predefined regarding the parameters of an environment, an RL agent is able to adjust its learning process continuously during acting. This is done by continuously updating the expected reward of an action (state-value) by means of a reward function. The agent learns automatically when it conducts an action that matches the reward function and can subsequently shape its behavior in order to increase future rewards. The feature that is most relevant for our analyses is that the reward function defines which action can count as a successful action and therefore as a learning progress.

Yet, one central problem consists in the initial reward as the RL agent has to exploit a state space randomly by trial and error in order to discover the first reward. To avoid a time-consuming random search the reward function has to be carefully designed. However, this limits the flexibility of the algorithm. In order to bypass an exclusively trial-and-error search or a complicated design process, coaching is implemented in the RL agent by adding an interface to the RL agent that allocates a feedback signal [17]. RL then allows for coaching in that the human trainer gives feedback, and the learning agent adjusts its reward function and its action rules according to the feedback. Thus, the behavior is not directly instructed or trained, but the robot modifies its behavior by itself. At the same time the reward function does not need to be designed in advance. This autonomous estimation of the reward function then complements the standard RL based on a direct interaction with the environment.

In the experiment an RL agent controls a robotic arm in order to swing up and keep an inverted pendulum balanced. While carrying out the task, the RL agent receives continuously feedback in terms of human subjective cues, i.e. positive or negative [29]. The agent has to interpret this feedback and adjusts the reward function and therefore its actions accordingly. Thus, learning the reward function is based on simple and abstract (binary) feedback that is delivered in social interaction. The feedback itself does not determine the reward function directly, but allows the robot to modify the latter based on an act of interpretation that consists in an estimation of the input's relevancy to its own behavior. This interpretation depends on two successive criteria. Firstly, in contingency or causality detection the "agent determines specific states [of its behavior] that motivated the trainer to give feedback" ([17], p. 5), i.e. the agent identifies the feedback's target behavior that depends on a certain time range and a subsequent time delay specifying the time between the action and the feedback. This identification of target behavior is, secondly, complemented by a consistency or error detection, i.e. checking to what extent a given evaluation corresponds "to current and previous feedback to a similar behavior" ([17], p. 5f.). If the feedback is inconsistent (contradictory), it is regarded as irrelevant and the reward function will not be updated. In short, after assigning the feedback to a previous action and verifying its consistency the evaluation function is updated and action rules modified accordingly. In this way the robot exhibits an internal and manipulable model of the trainer's evaluation instead of just executing correction commands. Hence, different kinds of feedback (coaching strategies) lead to different degrees of rates of learning and success.

The second experiment [14] was conducted by Gruebler, Berenz, and Suzuki. At first it has to be noted that we draw on the second experiment in order to exemplify the significance of social interaction while the behaviorand learning algorithm differs from the RL agent in the first experiment. However, due to the binary feedback in both experiments the results can be complemented in the subsequent investigation of subjective relations. Hence, the second experiment concerns the it allocation of feedback [14]. Human feedback is delivered as a cue based on a binary (positive or negative) signal that is interpreted as confirmation or correction. Continuous non-verbal social cues are used as instructive input to help a humanoid robot to modify its behavior. While the robot is conducting a task, the human coach gives continuous feedback by means of smiling or frowning. The facial expression was measured by a wearable device that recognizes these basal facial movements as expressions of confirmation (smile) and correction (frown) [15]. In this way a binary feedback resulted that enabled the robot to modify its behavior continuously whilst conducting a task. No further specification of the signal is necessary. In this way the robotic agent is open to human affective feedback in direct interaction. The cognitive and interactional implementations of both experiments can be complemented to that effect that a binary signal is sufficient to instruct a robot while at the same time this signal can be allocated in a way very natural for humans.

2.3 Experimental layouts and results

Experiments on coaching a robot based on human subjective feedback form the ground for an analysis of a subjective agent. Both experimental setups that were introduced in the previously, are cases of HRI. The RL agent of the first experiment [17] has been implemented in a simulated and a real robotic arm whose learning task consisted in swinging up and keeping an inverted pendulum balanced (see Fig. 1). Instead of predesigning the reward function, the human instructor assists the RL agent by observing its behavior and giving a binary (positive or negative) feedback. In the real and the simulational setup a "significant improvement compared to the conventional RL with the same reward function" ([17], p. 14) had been measured as the conventional RL completely failed to achieve the task. The simulational setup additionally showed that the RL agent reflects coaching strategies of different instructors in that one instructor failed to assist the RL agent as she gave too many negative feedbacks.

In the second experiment [14] a human instructor assisted a humanoid robot in a sorting game. The goal was to give red balls to the instructor and to throw green balls away. The affective feedback was detected by a facial expression reader [15]: smiling (positive) for confirmation of an action and frowning (negative) for correction (see Fig. 2 and 3). The robot successfully learned the desired task and was able to sort the last two balls without assistance. Furthermore, it proved that coaching by affective feedback leads to a significant improvement of HRI as the human instructor can act in a very natural (human-like) manner [14].

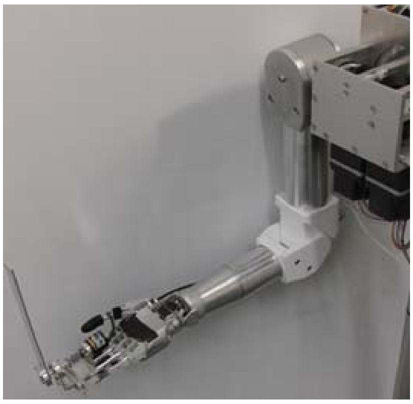

Figure 1. Robotic arm swinging up and keeping a pendulum balanced (figure taken from [17]).

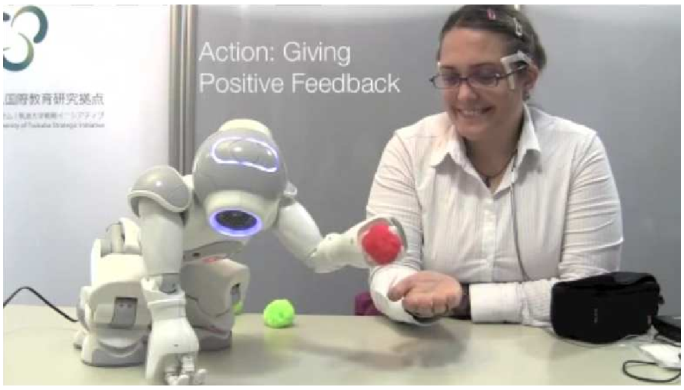

Figure 2. Interaction with positive feedback (figure made available by Anna Gruebler).

3 THE METHOD OF LEVELS OF ABSTRACTION

Floridi proposes the method of levels of abstraction to analyze a system at different epistemic levels (cf. [10], ch. 3). This method to analyze all kinds of systems is inspired by the so-called Formal methods, a technique of computer science that aims at modeling a computer system regarding the "initial statement of a customer's requirements, through system design, implementation, testing, debugging, maintenance, verification, and evaluation" ([30], p. 8). Floridi utilizes this approach to evaluate a system technically for an epistemic analysis. In the line of Kant's critical philosophy [20], he stresses the epistemological issue to consider "the conditions of possibility of the analysis (experience) of a particular system" ([10], p. 60). This recourse to the conditions of possibility of an analysis is crucial in order to avoid the mistake of analyzing a system independently of any specification of the analysis. These specifications, firstly, comprise the goal or purpose of an analysis. Furthermore, based on the general distinction that we can analyze a given system regarding its onotological levels of organization (LoO) and epistemological levels of explanation (LoE), levels of abstraction (LoA) serve to make explicit the ontological and epistemological commitments of the analysis. Thus, LoA guide the analysis teleologically towards a certain goal of interest.

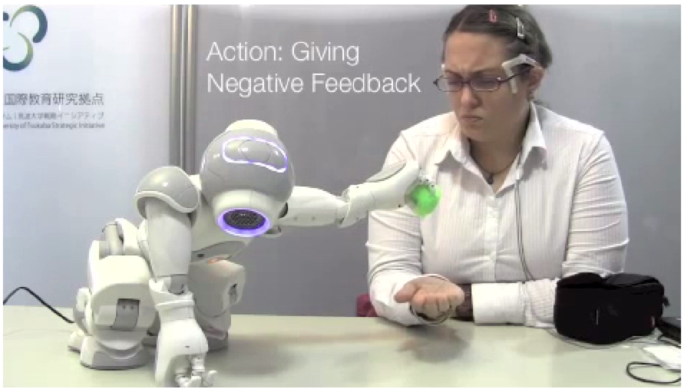

Figure 3. Interaction with negative feedback (figure made available by Anna Gruebler).

As an epistemic levelism, each LoA depends on a certain observation or interpretation of a system. Hence, the technical concepts of the method of levels of abstraction and their formal definitions mainly comprise typed variables and observables, defined as follows (following quotations are from [10], ch. 3.2):

1. "A *typed variable* is a uniquely named conceptual entitiy (the variable) and a set, called its type, consisting of all the values that the entity may take." (p. 48)
2. "An *observable* is an interpreted typed variable, that is, a typed variable together with a statement of what feature of the system under consideration it represents." (ibid.)
3. "A *level of abstraction (LoA)* is a finite but non-empty set of observables." (p. 52)

Different LoAs of a system are integrated in a Gradient of abstraction (GoA), i.e. a LoA allows to specifically model a system, whereas in a GoA we can switch between different LoAs. To facilitate such a leveled analysis of a system, certain relations on a LoA and between all LoAs of a GoA must hold. The LoA-specific constraint is defined in terms of behavior:

4. "the *behaviour* of a system, at a given LoA, is defined to consist of a predicate whose free variables are observables at that LoA. The substitutions of values for observables that make the predicate true are called the system behaviours. A moderated LoA is defined to consist of an LoA together with a behaviour at that LoA." (p. 53)

Based on moderated LoAs, the GoA is defined as follows:

5. "A *gradient of abstractions*, GoA, is defined to consist of a finite set $\{L_i \mid 0 \leq i < n\}$ of moderated LoAs L_i, a family of relations

$R_{i,j} \subseteq L_i L_j$, for $0 \leq i \neq j < n$, relating the observables of each pair L_i and L_j of distinct LoAs in such a way that:

(a) the relationships are inverse: for $i \neq j$, $R_{i,j}$ is the reverse of $R_{j,i}$

(b) the behaviour p_j at L_j is at least as strong as the translated behaviour $P_{R_{i,j}}(p_i)$." (p. 55)

The GoA applied in our analysis of the coaching RL agent will be a *nested* GoA, i.e. its "non-empty relations are those between L_i and L_{i+1}, for each $0 \leq i < n-1$, and moreover the reverse of each $R_{i,i+1}$ is a surjective function from the observables of L_{i+1} to those of L_i." (p. 56)

Observations at one LoA can generally be related to observations at another LoA, but there are different ways of relating LoAs. Most prominently are hierarchical GoAs that propose one detailed LoA that serves to explain the observations at a more abstract LoA. This is for example the case in neurophysiological models of cognitive abilities where the biochemical reactions form the basic LoA. Cognitive abilities are modeled at more abstract or higher levels so that the observables at a higher level (e.g. phenomenal experience) can be translated to observables at a lower level (neurophysiological reactions). Whereas a hierarchical GoA can imply a reductionist approach, we make use of a net of abstractions, i.e. it is not our goal to reduce mental abilities to computational processes. Hence, we do not follow an ontological approach in order to determine the nature of mental or subjective states. Instead, we follow a *functional* approach in order to make explicit the functional organization of the coaching RL agent's information space [10], ch. 3.4.7. Accordingly, different LoAs are related by simulation, i.e. one LoA simulates the behavior of another LoA. The simulation relation connects different LoAs by a mapping relation R that relates the observables of two LoAs mutually. Unlike a hierarchical GoA or even a reductionist model of explanation, there is no basic or foundational LoA that realizes other LoAs unidirectionally. Instead, one system (here the coaching RL agent) is considered in terms of different functional realizations that are mutually related by a simulation relation. In a nested GoA, for every single observable at one LoA, it can be shown how this observable behaves at another LoA. In this way, different LoAs can be connected and serve as *mutual* explanation of their behavior. According to this mutual explanation of behavior the GoA serves to link different epistemic LoAs.

Our analysis of the coaching RL agent is placed in the broader context of subjective computing that was used to solve a learning task (see section 2.2 and 2.3). More precisely, we want to determine to what extent the algorithmic implementation can be related to a mental description of the agent's behavior. As mental abilities presuppose a subject that acts mentally, our analysis concerns the relational structure of the agent's information space. By means of this relational analysis, firstly, the kind of relations that hold between the agent and its environment (relation to others) and within the agent (self-referentiality) can be made explicit. By means of this relational account, we can, secondly, decode mentalistic terms (observing, considering, interpreting) in terms of the other LoAs and finally determine to what extent the coaching RL agent can be accounted for as exhibiting mental and therefore subjective abilities.

This way of analyzing subjective abilities of a robotic agent might force the straightforward objection that mental or subjective abilities are haphazardly imposed on a system that does not really possess these. This objection is grounded in the method of LoA as every LoA is based on an abstraction of the system under consideration: an abstraction of certain features is only possible if certain other features are neglected. E.g., we can analyze a robotic system regarding the mechanics of its components, the programming framework, the costs of its production, or, as in our case, its relational structure. Taking into consideration one perspective onto a system, implies neglecting other possible perspectives. Regarding the coaching RL agent, we neglect any phenomenal description of its behavior as we focus on the relational structure. Accordingly, we may not expect to analyze the agent's (potentially) mental behavior in human-like psychological terms. In the face of full-blown human subjectivity, it has to be admitted that the ascription of mental or subjective states cannot be completely justified by means of a relational GoA as its observables are defined regarding the system under consideration (here the coaching RL agent). To compare with a human subject we would have to define observables that also cover human cognition. But a GoA is always related to a certain system, and our goal is not to compare the coaching RL agent with a human subject (a futile undertaking in that the robot is without any doubt less subjectively and cognitively equipped), but to investigate certain relational aspects that are constitutive for subjectivity in general. If these relational aspects of cognition are utilized for the design of an agent and this agent shows a significantly improved and more intelligent behavior than without these subjective features, the technical implementation of certain relational aspects of subjectivity may be interpreted as a confirmation for the underlying philosophical concept of subjectivity. The methodological presupposition that justifies this ascription of mental abilities in favor of relational subjectivity, is based on a constructionist or design approach in philosophy [10], p. 72, 76ff.: a theoretical concept is validated and, at its best verified, if it is possible to design and implement a technical system according to this concept. Or, as in our case, if a technical implementation is shown to utilize this concept successfully.[5]

4 LEVELED ANALYSIS OF THE COACHING RL AGENT

Based on the method of LoAs we defined four LoAs in order to analyze the coaching RL agent:

1. *Algorithmic level.* This level depends on the algorithm that is implemented in the coaching RL agent. Whereas the computational level is fully covered by the original experiment [17], we focus on the cognitive abilities that are facilitated by the algorithm.
2. *Functional level.* The basic algorithm instantiates certain functions and therefore enables the agent to fulfill certain computational tasks; accordingly the agent determines, compares, and processes given feedback.
3. *Relational level.* The agent's information space depends on different kinds of relations to given input. For the following analysis it will be crucial to distinguish between a straightforward determination by direct world-coupling and a self-determination by means of a social relation that allocates feedback.
4. *Mentalistic level.* This level comprises the mentalistic description of the agent's actions. The goal of this analysis is to investigate to what extent the algorithmic, functional, and relational level allow for a mentalistic and therefore subjective characterization of the coaching RL agent.

Before we go into a formal treatment in order to bring forward a nested GoA of the coaching RL agent, we offer an informal treatment

[5] The constructionist approach asks for a continuative justification that exceeds the scope of this paper; see [10] for further discussion.

of coaching a robot. This serves to make clear at which levels we analyze the agent and how we relate the cognitive and interactional capacities to the mentalistic description of the agent's behavior. Based on this informal and the subsequent formal treatment it will be possible to evaluate to what extent the ascription of mental abilities is justified.

4.1 Instantiating a subjective agent in social interaction

The task of coaching a robot offers an instructive way to study the behavior of an autonomous agent that interacts with humans. One special feature consists in the mutual exchange between the robotic agent and the human trainer. The agent is not only supposed to deliver a computational result as for instance in the case of search engines, but its actions provoke the trainer's feedback that itself serves the agent to modify its actions. Even if the exchange between robot and human does not take place on a linguistic level, the trainer's feedback is answered by the robot's behavior whereas the behavior provokes new feedback. To improve the learning abilities of the robotic agent a RL agent was complemented with a coaching function (see section 2.2). This functional follows two central purposes: By means of the feedback the RL agent can adjust the learning parameter (reward function) that defines the success of an action during the learning process. On the other hand the coaching function enables a human trainer to interact with a robotic agent in a very natural (i.e. affective) manner. The trainer just gives positive or negative feedback that is to be processed autonomously (interpreted) by the robotic agent. By allocating the feedback by means of a facial emotion reader [15] the mental workload for the human trainer decreases to a minimum level that does not differ significantly from a human-human interaction.

Our case study [17] is based on a robot arm platform (see Fig. 1). The robot has to solve the task of keeping a pendulum balanced. In order to accomplish this task the agent can modify the joints of its arm to handle the pendulum. But it has to learn how to modify its joints. In the coaching framework a human trainer gives a two digit feedback (positive or negative) while the agent is trying to keep the pendulum balanced. Accordingly, the robot must be able to process the feedback. The final goal is that the agent processes the feedback and adjusts its actions autonomously. As we deal with a robotic system we basically have to take into account the algorithmic implementation of the cognitive abilities required to process the feedback. So the basic LoA comprises of the algorithmic implementation.[6] Accordingly, at this level we should not conceive of an agent that acts, but of algorithmic processing. In our case study the robotic agent is able to react to feedback in a twofold manner. The algorithm enables the robot to determine which of its behavior refers to a feedback. This step of determining the feedback's target behavior (causality detection) is crucial for the processing of the feedback as the robot must be able to relate a feedback to its behavior. Even in human-human learning we know the common misunderstanding that the trainee sometimes allocates the feedback to a different behavior as the trainer aimed at. Furthermore, the algorithm allows the coaching RL agent to compare a feedback with previous feedbacks related to the same action. This test for consistency serves to identify contradictory feedback as an action cannot be conceived of as a successful action based on positive feedback when at the same time the action was evaluated negatively earlier. Again, the consistency of feedback is even crucial for human-human learning as a trainee can benefit from unambiguous feedback whereas contradictory feedback already presupposes a certain level of expertise if the trainee is supposed to profit in the same way as in the case of unambiguous feedback. Finally, when a feedback was assigned to a certain target-behavior and the feedback is consistent with previous feedbacks of this behavior, then the algorithm leads to a modification of the reward function and subsequently to adapted behavior. This final adaption of behavior can count as a successful learning process as the robot's behavior improved in order to accomplish the task to a higher degree than before the learning process.

In our example the learning process goes like this: When trying to balance the pendulum, the robotic arm platform starts with the initial posture of the pendulum as vertically downward. The robot decides how many degrees it moves its joint at every time step according to the current situation. Furthermore, it remembers the history of its actions. While balancing the pendulum, a human trainer gives positive or negative feedback. Via an interface this feedback is allocated as a reward value for the RL agent. Then every single feedback is processed according to the algorithm, i.e. the robot, firstly, determines the target behavior of a feedback. The target behavior of the feedback is the movement of the robot's joints within a certain time range. Hence, when the feedback is given from the trainer, the robot is able to estimate which of its actions the trainer actually evaluated by referring to a certain time range of the history of its actions. Whereas this time range, which the feedback refers to, can in principal also be learnt. In the experiment the time range was defined based on the measurement of human's delay of cognition. How much time passes before a human trainer gives feedback was measured: the results show that the minimum and maximum delay lies within 300 to 800[ms] (cf. [17], p. 11). According to this data, the coaching RL agent mapped a feedback to its behavior 300 to 800[ms] ago. After the determination of target behavior the agent, secondly, compares the feedback to previous feedbacks of the same behavior. If the previous acceptance or denial of this behavior is confirmed, the robot modifies its reward function accordingly. Based on this adjusted reward function the agent prefers the actions that were evaluated positively and changes its action rules. Thus, when the feedback confirmed a certain modification of the joints, then the agent will modify its behavior in order to move its joints according to the confirmed behavior. If, for example, the position of one joint within a certain range provoked positive feedback, then the robot will not exceed this range. Or if a certain joint angle provoked only negative feedback, the robot will not move this joint any more to this degree.[7]

Obviously, the previous description of the algorithmic level does not capture a mental or subjective ability. It entails the description of data processing and the transformation of data into modified behavior. But when we conceive this algorithmic processing at a functional level, we can take into account the functions instantiated by the algorithm. The functional description refers to the causal role of a component and specifies how an agent is empowered to act [4], [3]. The functional level allows to abstract from the algorithmic as computational processes and conceive the latter as cognitive functions of an agent. This shift of our investigation is crucial as on the algorithmic level there is strictly speaking no agent acting, but an algorithm is processing data. The fact that the algorithm enables an agent cannot be made explicit until we shift our attention to a functional level. Here it is that the computational reward value becomes a feedback as

[6] The study of a robotic system, or more generally, of an algorithm guarantuees that the system is controllable and implementable, i.e. we deal with a *white box* so that all parameters and internal operations can clearly be specified (cf. [13]).

[7] In the actual experiment, the ability of interpretation was limited to the extent that the robot could not process prevailing negative feedback.

a feedback is only possible in the mutual exchange of agents, i.e. between the human trainer and the robotic agent. The trainer primarily interacts with the robotic agent and not with the algorithm. Whereas in a strictly computational perspective one might say that the human trainer interacts with the algorithm, this does not make sense if we investigate the coaching process from a cognitive perspective. Cognitively speaking the computational reward is a feedback that has to be translated into a computational format. But again, the human trainer is not directly giving a computational reward value but an affective feedback [15]. Thus, the whole importance of the difference between function and algorithm lies in the transformation of an affective reaction (positive/smile or negative/frown) into a binary reward value. Or, correspondingly, i.e. seen from algorithm to function, in the empowerment of an agent to operate on affective feedback. Hence, in a functional perspective we can actually conceive of a robotic *agent* that receives feedback. Functionally speaking, it is an agent that determines target behavior, compares and finally processes feedback. We shifted from an algorithmic description of computational processes to a functional characterization of an agent.

Whereas we proceeded from algorithmic processing to the capacities of an agent, the functional characterization still does not allow for a mental or subjective description of the coaching RL agent. Certain functions can be instantiated by many different systems that are obviously far from being mental or subjective. A thermostat fulfills the function of adjusting temperature or a search engine ranks data according to some specified criteria. So we have to take into account a further LoA that helps to identify if and to what extent the coaching RL agent is supposed to act subjectively. This is the relational LoA that models the agent's relations to itself and others. When conceiving of mental abilities (implied in the use of mentalistic language), we expect an agent that acts autonomously and is not just responding automatically to some input data. Hence, the agent's relations to some given input is crucial for evaluating its behavior [16].

Based on a relational analysis we can distinguish between different kinds of relations between the agent and its environment. On the one hand the agent's behavior is forced by standard RL that is based on direct world coupling. In standard RL, the behavior gets automatically modified by environmental constraints. This modification depends on the reward function as the criteria which actions count as a success and which actions fail to accomplish the task. In fact, our example displays an extreme case as when the pendulum fell down no further adjustment or modification of behavior is possible. The task inevitably failed. But in more flexible tasks, e.g. as in the case of navigation, environmental constraints could force an agent to change its direction when it encounters an obstacle. The crucial point here is that the agent's relation to an input (the obstacle) is determined, i.e. the agent's behavior changes automatically without that the agent does have any control of this modification of its behavior. Furthermore, all modifications depend on the predefined reward function. In the case of the coaching RL agent, this way of direct world-coupling is complemented by an autonomous self-relation. The robotic arm not only reacts automatically to external events (here that the pendulum falls down). The agent is able to operate on the automatic learning process so that this process is not any more comletely determining the agent's behavior. Based on the algorithmic causality and error detection, or the functional capacity to determine target behavior and compare feedback respectively, the agent is able to process a binary feedback and decide by itself whether and to what extent its behavior should be modified. Both relations, the direct world-coupling and the interpretation of feedback, contribute to the agent's performance.

One might object that the robotic arm does not engage in full-blown decision making, but that is not the point here. Here it is crucial that the agent's behavior is significantly improved in that the final modification of the behavior is left to the agent itself. The agent operates autonomously on feedback and therefore relates autonomously to its own internal model of the trainer's evaluation. Thus, autonomous self-referentiality comprises that an agent operates on its own internal states whereas these operations do not completely underlie any external constraints [16]. The underlying concept of autonomy does not aim at complete self-determined behavior. Instead, autonomous behavior can be generated in opposition to determined behavior, i.e. the determination of the agent gets limited, or, correspndingly, the agent's autonomous capacity has to be constrained in order to bring forward successful behavior. The theorem of 'realization by means of restriction'[8] clarifies the role of social interaction. In our case, social interaction lies between the autonomous interpretation and direct world-coupling, i.e. it is a *partial* determination of the agent's information space as the agent is constrained by the feedback values, but is autonomous regarding their further processing. Due to the difficulties to define a suitable reward function a priori (see section 2.2), the coaching function and the feedback were introduced in order to assist the robot with updating the reward function. Accordingly, the subjective momentum of the agent's informational space depends on a mutual dependence of the autonomous self-relation and the social interaction: in order to evaluate its behavior autonomously the agent depends on a certain input (feedback) that confines his capacity to interpret the feedback to some reasonable options. Otherwise, the agent would have no criteria how to evaluate its actions, i.e. how to move its joints.

The autonomous and at the same time partially determined behavior lies at the ground of a subjective agent and serves to identify the final LoA. The coaching RL agent can be regarded as acting mentally in that it interprets the feedback based on an autonomous decision making: the agent *considers* contingency, *observes* the consistency of given feedback which results in an interpretation. Mental states of considering, observing, and interpreting that presuppose a subjective agent are based on the mutual relationship of autonomous self-referentiality and social interaction in that the straightforward determination of behavior by direct world-coupling is interrupted. We can call the agent's interpretation 'mental' or 'subjectiv' as this behavior is finally determined in the agent's information space by the agent itself and not primarily by some external constraints. The robot, being socially receptive for direct interaction with a human and its autonomous decision making, qualifies the coaching RL agent as a basically subjective agent. Again, one might object that this kind of subjectivity is less than what we usually ascribe to full-blown human subjects. But despite these obvious restrictions, the leveled analysis of the robotic agent offers us an account of subjectivity that does not rely on intractable phenomenal or psycgological states. We can instead follow the generation of a subjective agent from scratch. Furthermore, we are forced to include social interaction, which easily gets lost in phenomenal accounts. The main purpose of the following formal treatment lies in the need to make explicit the relations within and between every LoA, as subjectivity is here primarily seen under a relational viewpoint.

4.2 Nested GoA of the coaching RL agent

According to the previous stated method of levels of abstraction and the informal treatment, the RL agent is now to be analyzed formally

[8] See the chapter on schematism in [20], and [10].

at four LoAs. Each LoA comprises three observables (interface, interpretation, learning) with specific variables related to the observables. The relational LoA forms an exception, in that not the observables themselves but the relational structure of the agent's processing describes the behavior of this LoA. The following formalization does not depend on any specific mathematical standard but merely seeks to make clear the different levels of the agent's cognitive activity and especially the relations between the agent and the trainer's feedback at L_2.

The nested GoA is based on the following levels (L), comprising the observables interface, interpretation, learning, and corresponding variables:

- L_0: *algorithmic level*
  - Interface: reward value V
  - Interpretation: estimation of reward function E_F
  - Learning: updating reward function and action rules U
- L_1: functional level
  - Interface: feedback F
  - Interpretation: estimation of relevance E_R
  - Learning: processing feedback F_p
- L_2: relational level
  - Agent's self-referentiality: A_s
  - Agent's social relation (interaction): A_i
  - Direct world coupling (standard RL): A_d
- L_3: mentalistic level
  - Interface: social receptivity S
  - Interpretation: interpreting feedback F_i
  - Learning: reflecting feedback F_r

These observables and corresponding variables form a nested GoA of the coaching RL agent (see table 1). The GoA consists of four LoAs specified in the first column and beginning with the algorithmic level. Due to the epistemic foundation of LoAs, the epistemic regard according to which the coaching RL agent is interpreted is given at each level. Each LoA consists of three observables: interface, interpretation, and learning. On L_0 the system is analyzed regarding its algorithmic processing. This computational level is fully covered by the original experiment [17]. In the following analysis we therefore solely focus on those aspects concerning the instantiation of an information space which is computationally implemented as a continuous state-action-space: a certain signal, the reward value V is delivered by an interface and gets processed in the course of causality and error detection. In the course of interpretation, these processes of detection are regarded as an estimation of the reward function E_F. Learning at the algorithmic level consists of an updated reward function and correspondingly updated action rules U. The system's behavior can be specified by the use of the following predicates: V delivers a binary value corresponding to a positive (smile, V^+) or negative (frown, V^-) evaluation of the instructor. E_F delivers a value under or above the current reward function and leads in the first case to an update U of the reward function and the action rules so that U contains the updated and modified reward function that will result in adapted behavior.

At the subsequent LoAs these processes remain the same, but are analyzed differently. Considering the functionality of the algorithm at L_1, the algorithm enables an agent to fulfill certain computational tasks: the agent determines, compares, and processes given feedback. This functional mapping serves to identify the cognitive processes of the agent (cf. section 4.1) as follows: The algorithmic observables at L_0 are mapped to L_1 as follows: V functions as feedback and takes the values of 'confirmation' V^+ or 'correction' V^-, i.e. the function of the computational values consists in allocating positive or negative feedback. Hence, the meaning of the computational value V for the agent's behavior is identified by its function. The same counts for the estimation of the reward function E_R: E_F fulfills the cognitive function of specifying the feedback according its relevancy for the agent's behavior. E_R is the result of estimating V, i.e. $E_R = E_F(V)$. Finally the cognitive function of U is processing feedback by updating the reward function and the action rules in order to increase learnability, i.e. $F_p = U(E_F)$.

L_2 contains the crucial relational analysis. The agent's information space depends on two kinds of relations that hold between the previous observables and the agent's capacity to operate on them. On the one hand the agent underlies two determining relations to others: Based on the implementation of standard RL the coaching RL agent depends on external and automatic determination of behavior through direct world coupling A_d and a subsequent adaption to the environment. Secondly, in the course of social interaction A_i the trainer allocates feedback F. Whereas the feedback contains a fixed value (positive or negative) that cannot be altered by the agent, the further processing of the feedback is subject to an interpretation by the coaching RL agent that decides if its behavior gets modified. Hence the degree of determination of the agent's behavior decreases significantly in the course of social interaction. The subjective momentum of the agent's information space is generated by the second kind of relation, i.e. the agent's autonomous relation to its own internal model of the trainer's evaluation. The RL agent is able to modify the reward function and action rules autonomously and therefore indirectly its behavior. This relation to the incoming feedback is autonomous as the latter does not determine the agent necessarily or immediately as is the case with standard RL. Opposed to externally determined behavior as the result of A_d, subjectively modified behavior is instantiated by autonomous acts of interpretation by the agent itself. Thus, the agent's subjective information space depends on these simultaneous relations as can be made explicit by mapping the observables at L_1 onto L_2: According to standard RL the agent is determined directly through direct world coupling A_d. At the same time feedback is allocated by social interaction $A_i[F]$ that constrains the autonomous modification of the reward function: F is processed by E_R to F_p depending on the agent's own, i.e. subjective, interpretation of the feedback $A_s[E_R(F) \rightarrow F_p]$. The agent's autonomy consists in its ability to modify its own learning process by adjusting the reward function by itself. From a relational viewpoint, the agent's subjective determination of the reward function is constituted simultaneously with an objective determination of its behavior by direct world-coupling (see section 4.1).

The complete behavior of the coaching RL agent at L_2 depends on these parallel processes. Whereas determined behavior alone is not a special characteristic of *subjective behavior*, autonomous self-referentiality (A_s) and social interaction (A_i) are relevant for the final LoA, the mentalistic level. The subjective ability to modify the automatic learning process by autonomously processing feedback forms a necessary condition for subjective computing. At the same time autonomous self-referential behavior can only be effectively utilized in the course of social interaction as the agent has to learn how to modify its learning process. Hence, autonomous self-

Table 1. Nested GoA of the coaching RL agent.

LoA *Relations*	**Observables**		
L_0: algorithmic (algorithmic processing)	**Interface** reward value V	**Interpretation** causality detection → error detection, i.e. estimation of the reward function E_F	**Learning** updating reward function → updating action rules U
$R_{0,1}$: mapping algorithm to functions	$R_{0,1}(V, F)$ $F = V^+ \vee V^-$	$R_{0,1}[E_F, E_R]$ $E_R = E_F(V)$	$R_{0,1}(U, F_p)$ $F_p = U(E_F)$
L_1: functional (functions realized by the algorithm)	**Interface** feedback F	**Interpretation** agent determines target behavior (contingency) → compares feedbacks (consistency), i.e. specifies feedback by estimating its relevance E_R	**Learning** agent processes feedback F_p
$R_{1,2}$: mapping functions to relations	$R_{1,2}[(F, E_R, F_p), (A_s, A_i), A_d]$ $A_s[E_R(F) \rightarrow F_p] \wedge A_i[F] \wedge A_d$		
L_2: relational (relational structure of agent's processing)	**Self-referentiality** agent relates (based on autonomous acts of estimating) to its own internal model of the trainer's evaluation, i.e. an autonomous and subjective self-relation A_s	**Relations to other** *Social relation* trainer's evaluation of the agent's behavior (feedback F) is allocated in social interaction A_i	*Direct world coupling* determined and objective relation A_d based on direct world coupling (standard RL)
$R_{2,3}$: mapping relations to mental abilities	$R_{2,3}(A_s, S)$ $S = A_s(F)$	$R_{2,3}(A_s, F_i)$ $F_i = A_s(E_R)$	$R_{2,3}(A_s, F_r)$ $F_r = A_s(F_p)$
L_3: mentalistic (mental abilities)	**Interface** social receptivity S	**Interpretation** agent considers contingency, carefully observes consistency of given feedback, i.e. interprets feedback F_i	**Learning** agent learns autonomously, i.e. agent reflects feedback or differences of coaching strategies by its behavior F_r

referentiality and social interaction interdependently enable a subjective agent. Again, subjectivity here means that the robotic agent is able to modify an ongoing automatic process whereas this modification is externally supported (here by feedback) but is finally left to the agent's decision. Those subjective and interactional issues arise in scenarios where a robotic agent is supposed to adopt a task and to accomplish this task autonomously (e.g. driving assistance, search and rescue applications, or autonomous control in hybrid assistive limb [**?**]). But due to the difficulties of defining the robot's actions in advance or to define a suitable reward function a priori, social interaction (coaching) can be utilized in order to support the robot's autonomous modification of its behavior and therefore improve its learnability.

The relational structure and the instantiation of a subjective relation in the agent's information space finally allow for a mentalistic interpretation of the coaching RL agent at L_3. Usually, we ascribe acts like considering and reflecting to a full-blown subject. This is, obviously, not the case here. Full-blown subjectivity depends on further features like natural language and ethical addressability. But when taking into account the social interaction of the coaching RL agent, this agent acts as an autonomous counterpart of the human, i.e. the agent exhibits a sufficient level of autonomy that we can ascribe mental activity to it as follows: in operating on the instructor's input, i.e. autonomously relating to the feedback $A_s(F)$, the agent becomes socially receptive S in the course of interaction. The RL agent shows subjective behavior when individually and situation-dependently interpreting feedback $F_i = A_s(E_R)$ and correspondingly learning by updating the reward function and action rules according to its interpretation, i.e. autonomously processing or reflecting feedback $F_r = A_s(F_p)$. The social interaction between the trainer and the robotic agent is crucial for A_i and the mentalistic character of the RL agent's behavior as the feedback offers an additional input (binary cues) opposed to strict world-coupling in standard RL. The subjective momentum, based on autonomous self-referentiality, occurs as the RL agent's non-deterministic consideration of contingency and observation of consistency of feedback as well as in the subsequent reflection of differences of coaching strategies by means of more or less successful learning. There is no predefined reaction or development of the coaching RL agent's behavior, but subjective behavior due to the internal indeterminacy of the modification of the learning process. At the same time the agent's autonomous ability relies on social interaction that guides its ability to modify its learning process. Without this guidance the agent would not be able to execute its autonomous modification of the reward function as it has no information how and to what extent a modification might support to accomplish its task.

5 CONCLUSION

We wanted to investigate what it can mean for a robotic agent to behave subjectively. We approached this question by analyzing to what extent mental abilities can be ascribed to a robotic agent. In the course of analyzing a coaching RL agent at four LoAs we made explicit a relational level (L_2) that shows how mental abilities can be ascribed to the agent: the coaching RL agent behaves subjectively in that it is able to modify its own automatic learning processes by means of feedback that is allocated in social interaction. At the same time, the agent is still being determined by direct world-coupling. Hence, the relational level confirms a relational notion of subjectivity.

On the other hand, this result underlies a certain caveat in that the nested GoA of the coaching RL agent is based on an abstraction that focuses on the relational structure of the agent, i.e. we analyzed to what extent the agent's actions are self-referential and related to others as well as self-determined and externally determined. This relational account of the robot's information space does not cover a common psychologically or phenomenally based description of human-like cognitive processes as it is mainly decoupled from the concept of consciousness and linked to intelligence. From a relational viewpoint, consciousness is regarded as cognitive product. Hence, it is necessary to go back to a level of abstraction that does not presuppose any conscious states if conscious, or less difficult, mental abilities have to come into reach of an explanation. By modeling relational features of intelligence by means of a technical implementation, we gained an analysis of cognitive abilities that is fully tractable and implementable.

Based on a technical implementation that showed a significant improvement of an agent's behavior by means of the coaching function, it was relationally justified to conclude that the coaching RL agent acts subjectively as it makes effective use of autonomous self-referentiality and social interaction. The agent's subjectivity is generated in this course of action as the agent's self-determined behavior opposed to external determination by direct world-coupling. By means of this relational abstraction of the coaching RL agent, we can link the technical implementation with the conceptual foundation of subjectivity and subjective computing, respectively. With regard to the further development of subjective agents, the link of the technical and theoretical domain supports the improvement of subjective abilities. The theoretical framework of relational subjectivity can guide an extension of self-referential processing in order to allow the coaching RL agent to process ambiguous feedback. Another open question concerns social interaction in other modes than binary feedback. With regard to full-blown human subjectivity, the relational account does not exclude modeling more complex cognitive abilities as the use of natural language or ethical addressability. On the other hand, the theoretical framework of relational subjectivity is being modeled in the course of technical implementation. This allows us to test and verify a relational modeling of subjectivity.

ACKNOWLEDGEMENTS

We would like to thank the referees for their comments which helped improve this paper.

REFERENCES

[1] Cynthia Breazeal, 'Toward sociable robots', *Robotics and Autonomous Systems*, **42**(3-4), 167–175, (March 2003).

[2] Daniel Breazeale and Günter Zöller, *The system of ethics: according to the principles of the Wissenschaftslehre*, Cambridge University Press, Cambridge, UK; New York, 2005.

[3] Mark B. Couch, 'Causal role theories of functional explanation', *Internet Encyclopedia of Philosophy*, (2011).

[4] Robert Cummins, *The Nature of Psychological Explanation.*, MIT Press, Cambridge, Mass., 1983.

[5] Kerstin Dautenhahn, 'A paradigm shift in artificial intelligence: Why social intelligence matters in the design and development of robots with Human-Like intelligence', in *50 Years of Artificial Intelligence*, eds., Max Lungarella, Fumiya Iida, Josh Bongard, and Rolf Pfeifer, volume 4850 of *Lecture Notes in Computer Science*, 288–302, Springer Berlin/Heidelberg, (2007).

[6] Kerstin Dautenhahn, Alan H. Bond, Lola Canamero, and Bruce Edmonds, *Socially intelligent agents: creating relationships with computers and robots*, volume 3 of *Multiagent Systems, Artificial Societies, and Simulated Organizations*, Springer, Berlin/Heidelberg, May 2002.

[7] Kerstin Dautenhahn, Bernard Ogden, and Tom Quick, 'From embodied to socially embedded agents – implications for interaction-aware robots', *Cognitive Systems Research*, **3**(3), 397–428, (September 2002).

[8] Johann Gottlieb Fichte, 'Grundlage der gesamten wissenschaftslehre als handschrift für seine zuhörer (1794)', in *Fichte-Gesamtausgabe der Bayerischen Akademie der Wissenschaften. Bd. I, 2.*, eds., Reinhard Lauth and Hans Jacob, 251–451, Bad Cannstatt, (1962).

[9] Johann Gottlieb Fichte, 'Wissenschaftslehre 1811', in *Fichte-Gesamtausgabe der Bayerischen Akademie der Wissenschaften. Bd. II, 12.*, eds., Reinhard Lauth and Hans Jacob, 138–299, Bad Cannstatt, (1962).

[10] Luciano Floridi, *The Philosophy of Information*, Oxford University Press, Oxford, 2011.

[11] Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn, 'A survey of socially interactive robots', *Robotics and Autonomous Systems*, **42**(3-4), 143–166, (2003).

[12] Michael A. Goodrich and Alan C. Schultz, 'Human-robot interaction: a survey', *Found. Trends Hum.-Comput. Interact.*, **1**(3), 203–275, (January 2007).

[13] Gian Maria Greco, Gianluca Paronitti, Matteo Turilli, and Luciano Floridi, 'How to do philosophy informationally', in *Lecture Notes on Artificial Intelligence*, volume 3782, pp. 623–634, (2005).

[14] Anna Gruebler, Vincent Berenz, and Kenji Suzuki, 'Coaching robot behavior using continuous physiological affective feedback', in *2011 11th IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, pp. 466–471. IEEE, (October 2011).

[15] Anna Gruebler and Kenji Suzuki, 'Measurement of distal EMG signals using a wearable device for reading facial expressions', in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 4594–4597. IEEE, (September 2010).

[16] Patrick Grüneberg, *Projektives Bewusstsein. Th. Metzingers Selbstmodelltheorie und J.G. Fichtes Wissenschaftslehre*, Dissertation, Technische Universität Berlin, 2012.

[17] Masakazu Hirokawa and Kenji Suzuki, 'Coaching to enhance the online behavior learning of a robotic agent', in *Knowledge-Based and Intelligent Information and Engineering Systems*, eds., Rossitza Setchi, Ivan Jordanov, Robert J. Howlett, and Lakhmi C. Jain, volume 6276, 148–157, Springer, Berlin/Heidelberg, (2010).

[18] Axel Honneth, *The struggle for recognition: the moral grammar of social conflicts*, MIT Press, Cambridge, MA, 1996.

[19] Frank Jackson, 'Epiphenomenal qualia', *The Philosophical Quarterly*, **32**(127), 127–136, (April 1982).

[20] Immanuel Kant, *Kritik der reinen Vernunft (1781/1787).*, Hamburg, 1990.

[21] Norimasa Kobori, Kenji Suzuki, Pitoyo Hartono, and Shuji Hashimoto, 'Reinforcement learning with temperature distribution based on likelihood function', *Transactions of the Japanese Society for Artificial Intelligence*, **20**, 297–305, (2005).

[22] Thomas Metzinger, *Being no one. The Self-model Theory of Subjectivity.*, MIT Press, Cambridge, Mass., 2003.

[23] Anthony F. Morse, Carlos Herrera, Robert Clowes, Alberto Montebelli, and Tom Ziemke, 'The role of robotic modelling in cognitive science', *New Ideas in Psychology*, **29**(3), 312–324, (2011).

[24] Thomas Nagel, 'What is it like to be a bat?', *Philosophical Review*, **83**(4), 435–450, (1974).

[25] Momoko Nakatani, Kenji Suzuki, and Shuji Hashimoto, 'Subjective-evaluation oriented teaching scheme for a biped humanoid robot', *IEEE-RAS International Conference on Humanoid Robots*, (2003).

[26] Marcia Riley, Ales Ude, Christopher Atkeson, and Gordon Cheng, 'Coaching: An approach to efficiently and intuitively create humanoid robot behaviors', in *2006 6th IEEE-RAS International Conference on Humanoid Robots*, pp. 567–574. IEEE, (December 2006).

[27] Richard S. Sutton and Andrew G. Barto, *Reinforcement learning: an introduction*, MIT Press, Cambridge, Mass., 1998.

[28] Andrea L. Thomaz and Cynthia Breazeal, 'Teachable robots: Understanding human teaching behavior to build more effective robot learners', *Artificial Intelligence*, **172**(6-7), 716–737, (April 2008).

[29] Andrea Lockerd Thomaz. Socially guided machine learning. http://dspace.mit.edu/handle/1721.1/36160, 2006. Thesis (Ph. D.)–Massachusetts Institute of Technology, School of Architecture and Planning, Program in Media Arts and Sciences, 2006.

[30] Jeannette M. Wing, 'A specifier's introduction to formal methods', *IEEE Computer*, **23**(9), 8–24, (September 1990).