Towards Self-Explaining Social Robots: Verbal Explanation Strategies for a Needs-Based Architecture

Workshop on Cognitive Architectures for Human–Robot Interaction

Sonja Stange
Bielefeld University
sstange@techfak.uni-bielefeld.de

Hendrik Buschmeier
Bielefeld University
hbuschme@techfak.uni-bielefeld.de

Teena Hassan
Bielefeld University
thassan@techfak.uni-bielefeld.de

Christopher Ritter
Bielefeld University
critter@techfak.uni-bielefeld.de

Stefan Kopp
Bielefeld University
skopp@techfak.uni-bielefeld.de

ABSTRACT

In order to establish long-term relationships with users, social companion robots and their behaviors need to be comprehensible. Purely reactive behavior such as answering questions or following commands can be readily interpreted by users. However, the robot's proactive behaviors, included in order to increase liveliness and improve the user experience, often raise a need for explanation. In this paper, we provide a concept to produce accessible "why-explanations" for the goal-directed behavior an autonomous, lively robot might produce. To this end we present an architecture that provides reasons for behaviors in terms of comprehensible needs and strategies of the robot, and we propose a model for generating different kinds of explanations.

KEYWORDS

social robots; transparency; explanation; self-explanation; behavior architecture

1 INTRODUCTION

Enabling intuitive interaction between humans and robots is a primary objective in research on human–robot interaction and social robotics. Part of this is to provide users with an appropriate understanding of the robot's behavior as needed, e.g., to personalize or teach the robot through direct feedback or instructions (cf. fig. 1). Recent work has started to look at how users interpret and understand the behavior of autonomous social robots. One increasingly adopted view is that, from the perspective of the user, a robot's behavior needs to be grounded in some form of intents [31] that provide comprehensible *reasons* [25] for it. Users, however, often build interpretations of robot behavior that are prone to uncertainty, misunderstandings, or unwarranted attribution (e.g., due to anthropomorphization of the robot). A suitable understanding thus often requires additional information provided through some form of explanation (cf. fig. 1). Here we ask how a social robot itself can be enabled to provide understandable explanations of its behavior, and how those should be tailored for the different kinds of behavior it can exhibit.

Social companion robots have often been designed based on concepts of internal drives or needs. Especially animal-like social robots such as Sony's 'AIBO' dog [8] evoke close social relationships [13] and are designed to have qualities such as emotions, personality, or

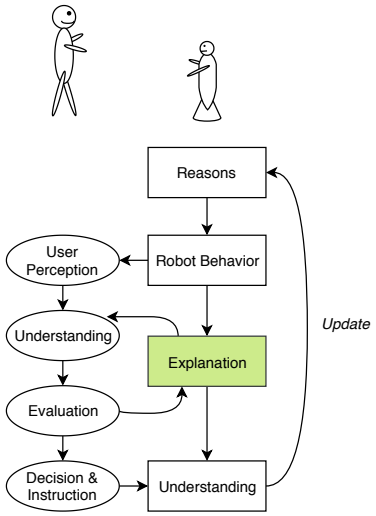

Figure 1: Explanations are an integral part of a comprehensible interaction and its coordination between robot (right) and human user (left).

attachment in order to become more similar to pet companions [16]. In this view, robots should *not* be perfectly obedient and without fault – a certain level of unpredictability and imperfectness would make robots more alive. However, while reactive behaviors may be intuitively interpreted by users in terms of stimulus–response patterns, the proactive behavior of a robot, initiated to, e.g., increase liveliness or propel self-driven learning, is often less transparent. We adopt the view that a robot needs to be capable of explaining its behavior in terms of *reasons*, in order for the user to understand origin and relevance of a behavior and to evaluate it accordingly. This self-explanation ability must render the robot's behavior transparent to the user, reducing her uncertainty and thus increasing trust towards the robot [19, 21, 28]. Further, behavior explanations should enable the user to evaluate the appropriateness of the robot's behavior under given circumstances and to provide feedback that allows for personalizing the human–robot interaction [20]. To that end, the user must be able to give informative feedback that

can be linked back to the robot-internal causal history of the action in order to adjust behavior generation mechanisms.

The research presented here aims at developing a lively social companion robot that is in "social resonance" [17, 27] with its user and elevates the user's well-being. In order to increase user trust in the robot and to enable learning from user feedback, our goal is to generate verbal explanations of robot behavior that are understandable by humans and give them the opportunity to make informed behavior evaluations, with the long-term objective of enabling users to teach the robot about their individual preferences and thus personalize the robot's behavior. The main question addressed in this paper, as a first step in this direction, is how the robot's action architecture needs to be structured to become "explainable", and how verbal explanations can be generated in this architecture to make the companion robot's behavior sufficiently transparent.

The paper is organized as follows: In section 2, background information and related work on models of behavior explanations in humans and robots are discussed. In section 3, a generic needs-based architecture for action selection in a social robot is outlined. Based on this architecture, section 4 proposes a model for generating different kinds of explanations, along with a discussion of concrete examples in a scenario for the environment of this social companion robot. Finally, in section 5, advances for refining explanations are put up for discussion and an outlook on future work is given.

2 BACKGROUND AND RELATED WORK

The question of how to explain the decisions or actions taken by autonomous AI systems has received tremendous attention in recent years. Besold and Uckelman [3] provide desiderata and guidelines for generating what they call *practical explanations* for decisions of artificial system: communicative effectiveness, accuracy sufficiency, truth sufficiency, and epistemic satisfaction (of the addressee). Their main focus lies on the epistemic dimension, strongly emphasizing that explanations vary by context and specifically the knowledge of the person requesting the explanation [3]. For the most part, this work on 'explainable AI' is aimed at increasing the transparency of hardly interpretable (black-box) models based on machine learning techniques, whereas we are concerned here with explaining the behavior of a social companions robot that rests on a complex perception–action architecture.

A large body of work has been directed to developing lively social robots. Animal behavior and cognition can be a source of inspiration for robots that embody individual characteristics and display varied, not completely transparent and predictable behavior. Both, the social robot 'Kismet' [4] as well as the aforementioned robot 'AIBO' [1], for example, employ control architectures that are loosely based on ethological principles and theories. Central to these architectures are components simulating a motivational system that models the robot's *needs* [26] as a set of internal variables that dynamically change, given the robot's actions and external stimuli. Depending on whether a certain need is fulfilled or not, the robot develops a corresponding *drive* [12] (e.g., social, stimulation, security, and fatigue in 'Kismet' [4]) that steers its behavior.

The motivational systems in 'Kismet' and 'AIBO' were designed specifically for these social robots with the goal to enable learning or increase entertainment-capabilities, respectively. More general cognitive computational models with motivational systems have been developed as well. The computational cognitive architecture 'CLARION' has recently been extended with a motivational representation [29] and the 'PSI'-model [7] specifically models motivation, emotion, and cognition. When combined with motivation-based action-selection mechanisms, as discussed in, e.g., [30], these cognitive architectures could – in principle – serve as a rich foundation for needs-based behavior-generation models for social robots.

It is generally acknowledged that a social robot should be perceived as intentional for users to establish a social connection with it [5, 31]. That is, humans should be able and willing to attribute beliefs, desires, or intentions to the robot. Yet, this attribution is not obvious if the behavior, e.g., derives from dynamic internal needs designed to simulate vivid, natural behavior. Many researchers have thus argued that such a robot should be equipped with the ability to explain its behavior in order to reduce the user's uncertainty and allow for a transparent and trusting interaction [14, 22, 28].

However, previous work on self-explaining intentional agents has mostly addressed training sessions, in which explanations are given primarily to enable more accurate task imitation [10, 11] or to educate the users [14, 15]. Harbers et al. [11] conducted a study revealing that users' explanations of an agent's behavior can be mapped to mental categories such as beliefs, desires/goals, and intentions (BDI). This suggests that in order to be explainable similar to human behavior, the agents should be designed according to BDI principles [10]. Further, they discovered that users prefer short (one to two elements), yet detailed explanations consisting of higher mental concepts. Kaptein et al. [14, 15] support the idea of BDI-based behavior explanation, but also emphasize the importance of personalizing explanations to the user.

An approach to explain agent behavior in terms of BDI principles is roughly in line with folk psychological studies on how humans come up with explanations for an observed behavior [23, 24]. According to the framework proposed by Malle et al. (cf. fig. 3A), one first differentiates between unintentional and intentional behavior. Intentional behavior is further explained in terms of *reasons*, concerning the agent's *desires* for an outcome (also referred to as *goals* or *aims*) as well as its *beliefs* that this specific action leads to a desired outcome based on a set of broad knowledge, hunches, and assessments. Further, there are various causal factors that can lead to the agent's reasons, such as its personality, culture, or the immediate context. Those are grouped under the term *causal history of reasons*, constituting a third type of explanation. These three factors lead to an agent's *intention* to perform a specific action, which is in turn contingent on personal (e.g., skill) and situational *enabling factors*. Such enabling factors are, for instance, used to explain unsuccessful attempts at an intended action or successful performance of an unlikely action [24].

The next section describes a concept for a needs-based architecture that makes similar explanations possible by opening up the behavior generation process to the user. Section 4 then describes how the robot can explain its behavior originating in this architecture, by using the four Mallean explanation types for intentional behavior (causal history of reasons, desire reasons, belief reasons, and enabling factors) [24].

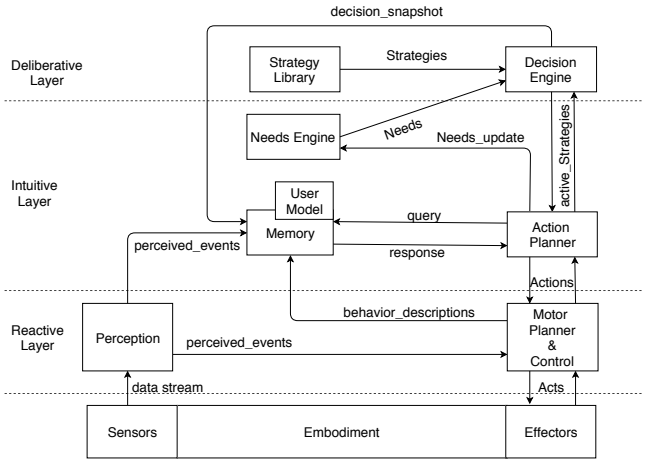

Figure 2: Architecture for a social companion robot that supports behavior explanation.

3 ARCHITECTURE

As stated before, our social robot has three requirements we like to see realized:

(1) liveliness through intrinsically motivated behavior,
(2) adaptivity through learning from user feedback, and
(3) explainability through verbalization of behavior motivating factors.

To achieve natural and lively behavior for our social companion robot, we base it on an architecture (cf. fig. 2) that integrates the necessary components for social robots to interact autonomously with their physical and social environment, and mechanisms enabling a robot to explain this behavior. The architecture is organized into three horizontal layers: reactive, intuitive, and deliberative. This allows modeling a closed body–mind loop via feedback loops that operate at different speed and cognitive complexity across the different layers. Such feedback loops can also be used to realize behavioral adaptation communicated via a social feedback loop with the user (requirement 2).

The *Reactive Layer* currently contains a base component for multimodal perception providing auditory and visual input from the robot's environment. A complementing *Motor Planner and Control* component synchronizes and executes low-level multimodal behaviors (gaze, facial expression, locomotion, speech), which we refer to as *Acts*. The *Acts* are considered voluntary if triggered by deliberative decisions (*Deliberative Layer*) or involuntary if they are the result of automated reactive behavior patterns in response to events from *Perception*. The *Acts* can currently be seen as represented by the elements defined in the Behavior Markup Language (BML) specification [2, 18].

The *Intuitive Layer* is in charge of creating behaviors that consist of learned or automatized reactions to higher-level perceived events. It is closely connected to a *Deliberative Layer*, which is responsible for strategic, goal-based behavior. One central component of the architecture at the interface of both levels is the *Needs Engine*, which models active *Needs* the robot is considered to have (compare, e.g., Dörner and Güss [7] or Maslow [26]). That is, besides reacting to its environment or user requests, the robot would have own needs that fluctuate depending on internal and external events. The *Needs Engine* component manages all active *Needs*, which we consider to create a physiological state – perceived as unpleasant – when unsatisfied [7]. Such valenced affect motivates the robot to act, although the agent can decide to suppress this urge, if necessary (see requirement 2). For a starting point, we identified two basic needs from our use cases: a need for physical integrity (relaxation) and a social need for affiliation (social contact with the user).

The *Decision Engine* examines the instantaneous *Needs-State* to identify the most pressing *Needs*. A *Strategy* is planned or chosen from a *Strategy Library* to satisfy the identified *Needs*, or a combination thereof. Selection criteria include the expected utility and the applicability of a strategy – verified with the help of *Perception* and *Memory*. The *Needs* can change over time, also depending on the results of executed *Strategies*. Robot *Needs* may be suppressed or have their impact adjusted to comply with user instructions. The *Needs Engine* integrates these updates received from different sources and regulates the *Needs-State*.

The *Action Planner* details and executes the selected *Strategy*. This component uses a dynamic planning approach to form a plan of *Actions*, to execute the selected *Strategy*. The goal of a plan is to perform activity that satisfies the activity-motivating need. For example, the need for relaxation is satisfied by performing the 'Battery Charging' *Action* for a sufficient amount of time; a need for socialization, on the contrary, is satisfied by any activity that involves interaction with the user. The *Action Planner* also reports progress updates to the *Needs* module to communicate *Needs*-based rewards from successfully performed *Actions*. Finally, *Actions* are mapped to *Acts* by the *Motor Planner and Control* component and communicated to the *Motor Layer* on the embodiment for execution. This of course is only applied if an *Action* has a physical manifestation or requires direct interaction with the environment.

Explanation (see requirement 3) requires the capability to relate internal motivational factors (*Needs*, *Strategies*, *Action-Plans*) with externally observable user behavior, and memories of past *Actions* and *Acts* performed and experienced by the robot. The latter aspect requires the architecture to provide memorizing and memory retrieval capabilities. This is realized through the *Memory* component. Percepts from the environment are stored in the *Memory*, which integrates new information into different cognitive representations that support spatial, semantic, episodic, and causal reasoning. The *Memory* is also used to maintain a Theory of Mind about the user – a mental *User-Model*.

Feedback from executed behaviors is stored in the memory as incremental snapshots of a behavior episode. The episode, along with the links to the initiating events and *Needs-State*, can be created in a BML-like fashion from performed *Acts*. The links especially allow the robot to identify whether a behavioral episode was just a reactive response to a perceived stimulus or a result of a deliberative decision to fulfill one of its own *Needs* (cf. requirement 3).

4 EXPLANATION MODEL

Explanations are highly complex and can be used with a lot of different possible functions and goals. For example, explanations could be produced to describe or clarify an action or event along

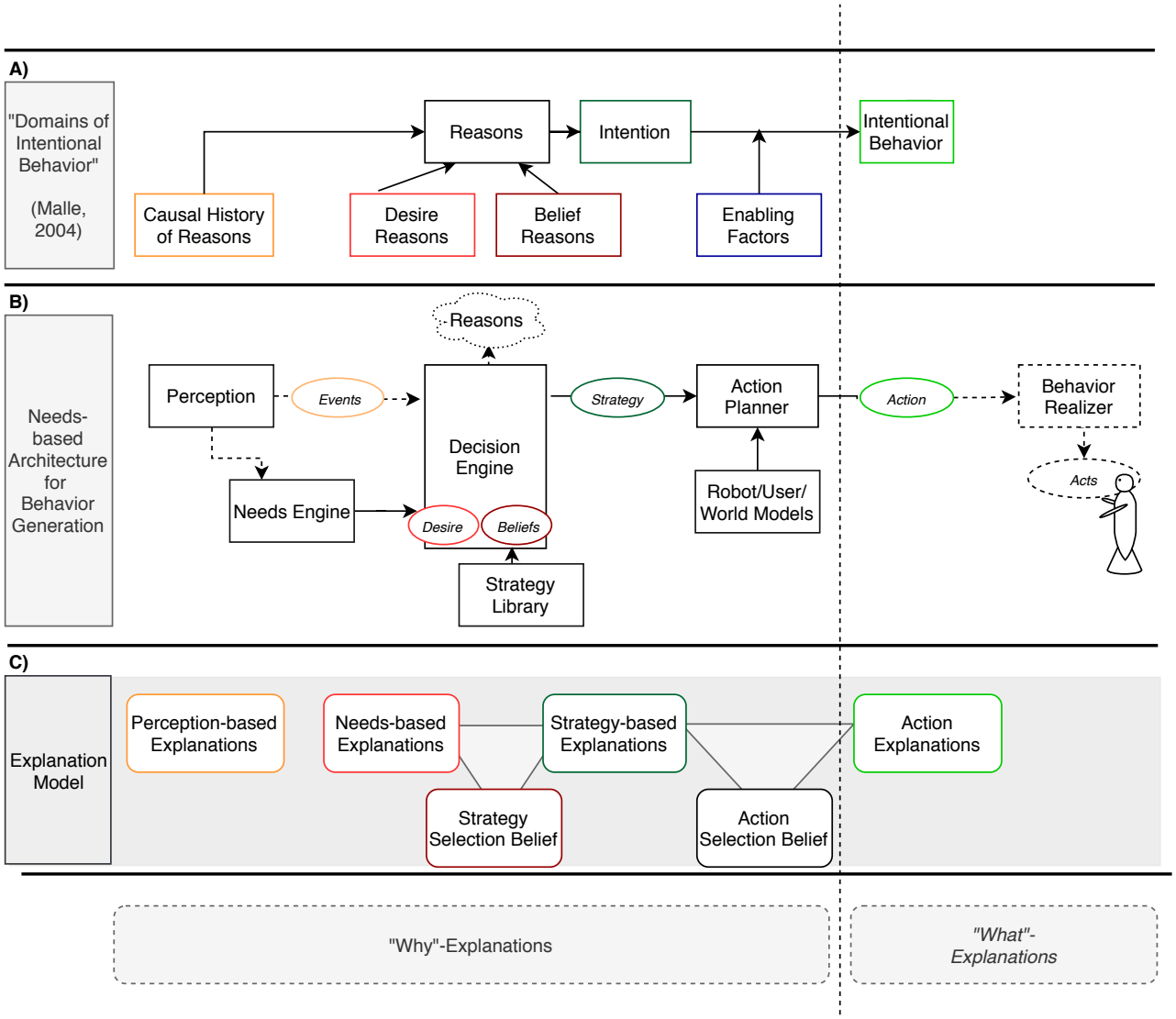

Figure 3: Different kinds of explanations (bottom) that a robot can generate for behavior originating in its needs-based architecture (middle) and based on the domains of intentional behavior found in human explanation (top; [24, p. 91]).

with its features ('what'-explanations). On the other hand, explanations could be meant to convey the underlying, hidden reasons for an action or event ('why'-explanations), possibly also to justify its occurrence. Further, explanations could aim at enabling the addressee to perform an action ('how'-explanations, such as detailed instructions). In all of these cases, explanations may even refer to a state of affairs or an action that has *not* taken place (e.g., contrastive or counterfactual explanations).

In the present work we concentrate on 'why'-explanations for 'social actions' that a companion robot is currently performing while being observed by a user. That is, we focus on the answers that the robot should give to a question like "Why are you doing this?". To this end, we need an explanation model that builds on the behavior production process in the needs-based architecture described in section 3, and that links it to folk-conceptual explanation strategies that humans are found to employ (cf. section 2). The underlying assumption is that people will better understand and prefer the explanations given by a social companion robot if they are similar to the explanations humans would give or come up with themselves [10].

4.1 Example Scenario

For the sake of clarity, we will use an example scenario to define and discuss the different possible types of explanation the robot should be able to give.

Imagine a mobile social companion robot, based on the previously described architecture, that is located in a private home. The

robot has different internal needs, including one for social contact and one for physical integrity (relaxation). Based on perceived events and its current needs, the robot can autonomously decide on which strategies may lead to a specific outcome and hence should be pursued (cf. fig. 3B). In our case, the robot is equipped with different strategies for satisfying its needs, e.g., for decreasing its loneliness it can establish social contact by either talking to a human or by establishing eye contact. Those general strategies comprise single actions as determined by the *Action Planner*, like driving along a particular path to avoid collisions with objects in the surrounding. In the particular situation considered here, the robot is driving along a peculiar path through the living room, while the user is sitting on the couch. The user now asks the robot to explain: "Why are you driving around here?".

4.2 Explanation Strategies

The answers to this question are plentiful and can vary broadly in range of quantity and quality. The presented architecture, however, offers a structured repertoire of possible explanations for the robot's behavior. Imagine the robot's behavior was generated as follows: The robot perceived the user sitting on the couch. Since its need for social contact was high and it believed that driving closer to the user may lead to the user talking to it, it chose the strategy of moving towards the user. In order to explain this action to the user, the robot can now choose from four different explanation strategies shown in fig. 3C:

- A *perception-based explanation* would inform the user on the input stimuli: "Because I saw you sitting there."
- Another simple, but rather obvious and thus unsatisfying explanation could be the *action explanation*: "I am moving closer to you".
- A *strategy-based explanation* would convey the currently pursued strategy, e.g.: "I wanted to get in contact with you".
- Finally, a *needs-based explanation* would be generated based on the robot's needs status that led to this strategy/action, e.g.: "I was lonely".

These explanations can either be used separately or combined in a more complex explanation strategy that links two (or possibly more) components to increase the traceability of the behavior generation process. The latter means to reveal the robot's decision-making process along with its beliefs about the relation between needs (and thus its desire to satisfy the most pressing needs) and the strategy that best addresses this needs distribution.

The necessity for this sort of complex behavior explanations can best be seen when looking at the following example: As before, the robot is driving through the apartment while the user is sitting on the couch. This time, however, the robot's behavior originates in its perception that outside the sun is shining, and that the *strategy* of seeking the sun may enable it to recharge its battery using solar energy (*belief*), which would satisfy its *need* of energy. It thus performs the *action* of driving to the balcony. Note that our concept of belief is broadly formulated, also entailing all the knowledge the robot has about the connections between different parts of its reasoning process.

In this scenario, basic explanations like "I want to recharge" or "I am seeking the sunlight" can be given as in the previous example. Yet, supposing that the user does not know that its robot is also rechargeable with solar energy (since it is a new robot or a new feature), a *needs-based explanation* would be rather irritating, since the user would not share the robot's belief about the sun leading to an increase in the energy level. A more elaborate and useful explanation would thus be: "I am seeking the sunlight, *because* I know that I can charge my battery with solar energy' (need + strategy selection belief), or "I am driving to the balcony, *because* I know that I only can get sunlight outside" (action + action selection belief).

5 DISCUSSION AND OUTLOOK

In this paper we have presented first steps towards a concept of how verbal explanations for behavior created in a needs-based architecture could be automatically generated, with the goal of elevating the transparency and thus reducing users' uncertainty when interacting with a social companion robot. In the long term, this should enable the user to give feedback and instructions that are in turn comprehensible to the robot, i.e., designed to target the same categories the robot uses in structuring and explaining its behavior and thus leading to concrete adjustments to actions, strategies, or needs at the different levels of the decision process.

Obviously, giving understandable and helpful verbal explanations in dialogue with a user goes beyond the conceptualization of behavior explanations such as the model laid out here. The style of verbal explanations needs to be chosen wisely and dynamically adapted to the explanation receiver's 'epistemic longing' [3] and personal attributes, such as their age or cultural background [14]. Malle [24] identifies three psychological determinants of explanatory choices, referring to the agent's behavior attributes, pragmatic goals, and information resources. Furthermore, explanations can be adapted linguistically, e.g., in their degree of politeness [6], which implies that they should change over the course of the human–robot relationship and depending on how 'delicate' the behavior to be explained is. Domain influences, such as expert vs. non-expert, also play a role in users' preferred explanation styles [9]. Different explanation types – such as belief-based (communicating the information on why the actor chose a certain action over another) or goal-based (communicating the actor's desired outcome) – are, for example, preferred for different actions [9] and by different age groups [14]. Similiarly, explanations enriched by simulated emotions could be beneficial for specific user groups [15].

Finally, appropriateness of the timing and complexity of explanations will need to be addressed: while many and highly elaborate behavior explanations might be preferred and needed at the onset of a human–robot relationship, choices need to be reconsidered when familiarity evolves, eventually limiting the necessity for behavior explanations to only a few occasions, for instance, when salient or unexpected behavior occurs.

ACKNOWLEDGMENTS

This research was supported by the German Federal Ministry of Education and Research (BMBF) in the project 'VIVA' (FKZ 16SV7959). The authors would like to thank the VIVA Team with special thanks to Claude Toussaint and Gheorghe Lisca.

REFERENCES

[1] Ronald C. Arkin, Masahiro Fujita, Tsuyoshi Takagi, and Rika Hasegawa. 2003. An ethological and emotional basis for human–robot interaction. *Robotics and Autonomous Systems* 42 (2003), 191–201. https://doi.org/10.1016/s0921-8890(02)00375-5

[2] Behavior Markup Language Committee. 2011. *The BML 1.0 Standard.* Retrieved from http://www.mindmakers.org/projects/bml-1-0/.

[3] Tarek R. Besold and Sara L. Uckelman. 2018. The what, the why, and the how of artificial explanations in automated decision-making. *CoRR* (2018). arXiv:1808.07074

[4] Cynthia Breazeal. 1998. A motivational system for regulating Human–Robot Interaction. In *Proceedings of the 15th National Conference on Artificial Intelligence*. Madison, WI, USA, 54–61.

[5] Cynthia Breazeal and Brian Scassellati. 1999. How to build robots that make friends and influence people. In *Proceedings of the 1999 IEEE/RSJ International Conference on Intelligent Robots and Systems*. Kyongju, South Korea, 858–863. https://doi.org/10.1109/IROS.1999.812787

[6] Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Usage.* Cambridge University Press, Cambridge, UK.

[7] Dietrich Dörner and C. Dominik Güss. 2013. PSI: A computational architecture of cognition, motivation, and emotion. *Review of General Psychology* 17 (2013), 297–317. https://doi.org/10.1037/a0032947

[8] Masahiro Fujita. 2001. AIBO: Toward the era of digital creatures. *The International Journal of Robotics Research* 20 (2001), 781–794. https://doi.org/10.1177/02783640122068092

[9] Maaike Harbers, Joost Broekens, Karel van den Bosch, and John-Jules Meyer. 2010. Guidelines for developing explainable cognitive models. In *Proceedings of the 10th International Conference on Cognitive Modeling*. Philadelphia, PA, USA, 85–90.

[10] Maaike Harbers, Karel van den Bosch, and John-Jules Meyer. 2010. Design and evaluation of explainable BDI agents. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. Toronto, Canada, 125–132. https://doi.org/10.1109/WI-IAT.2010.115

[11] Maaike Harbers, Karel van den Bosch, and John-Jules Ch. Meyer. 2009. A study into preferred explanations of virtual agent behavior. In *Proceedings of the 9th International Conference on Intelligent Virtual Agents*. Amsterdam, The Netherlands, 132–145. https://doi.org/10.1007/978-3-642-04380-2_17

[12] Robert Aubrey Hinde. 1956. Ethological models and the concept of 'drive'. *The British Journal for the Philosophy of Science* 6 (1956), 321–331.

[13] Peter H. Kahn, Nathan G. Freier, Batya Friedman, Rachel L. Severson, and Erika N. Feldman. 2004. Social and moral relationships with robotic others?. In *Proceedings of the 13th IEEE International Workshop on Robot and Human Interactive Communication*. Kurashiki, Japan, 545–550. https://doi.org/10.1109/ROMAN.2004.1374819

[14] Frank Kaptein, Joost Broekens, Koen Hindriks, and Mark Neerincx. 2017. Personalised self-explanation by robots: The role of goals versus beliefs in robot-action explanation for children and adults. In *26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. Lisbon, Portugal, 676–682. https://doi.org/10.1109/ROMAN.2017.8172376

[15] Frank Kaptein, Joost Broekens, Koen Hindriks, and Mark Neerincx. 2017. Self-explanations of a cognitive agent by citing goals and emotions. In *Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), 2017*. San Antonio, TX, USA, 81–82. https://doi.org/10.1109/ACIIW.2017.8272592

[16] Veronika Konok, Beáta Korcsok, Ádám Miklósi, and Márta Gácsi. 2018. Should we love robots?—The most liked qualities of companion dogs and how they can be implemented in social robots. *Computers in Human Behavior* 80 (2018), 132–142. https://doi.org/10.1016/j.chb.2017.11.002

[17] Stefan Kopp. 2010. Social resonance and embodied coordination in face-to-face conversation with artificial interlocutors. *Speech Communication* 52 (2010), 587–597. https://doi.org/10.1016/j.specom.2010.02.007

[18] Stefan Kopp, Brigitte Krenn, Stacy Marsella, Andrew N. Marshall, Catherine Pelachaud, Hannes Pirker, Kristinn Rúnar Thórisson, and Hannes Vilhjálmsson. 2006. Towards a common framework for multimodal generation: The Behavior Markup Language. In *Proceedings of the 6th International Conference on Intelligent Virtual Agents*. Marina del Rey, CA, USA, 205–217. https://doi.org/10.1007/11821830_17

[19] Dung N. Lam and K. Suzanne Barber. 2005. Comprehending agent software. In *Proceedings of the 4th International Joint Conference on Autonomous Agents and MultiAgent Systems*. Utrecht, The Netherlands, 586–593. https://doi.org/10.1145/1082473.1082562

[20] Iolanda Leite, Carlos Martinho, and Ana Paiva. 2013. Social robots for long-term interaction: A survey. *International Journal of Social Robotics* 5 (2013), 291–308. https://doi.org/10.1007/s12369-013-0178-y

[21] Joseph B. Lyons, Garrett G. Sadler, Kolina Koltai, Henri Battiste, Nhut T. Ho, Lauren C. Hoffmann, David Smith, Walter Johnson, and Robert Shively. 2017. Shaping trust through transparent design: Theoretical and experimental guidelines. In *Advances in Human Factors in Robots and Unmanned Systems*, Pamela Savage-Knepshield and Jessie Chen (Eds.). Vol. 499. Springer, Basel, Switzerland, 127–136.

[22] Bertram Malle and Matthias Scheutz. 2018. Learning how to behave. Moral competence for social robots. In *Handbuch Maschinenethik*, Oliver Bendel (Ed.). Springer, Wiesbaden, Germany, 1–24. https://doi.org/10.1007/978-3-658-17484-2_17-1

[23] Bertram F. Malle. 1999. How people explain behavior: A new theoretical framework. *Personality and Social Psychology Review* 3 (1999), 23–48. https://doi.org/10.1207/s15327957pspr0301_2

[24] Bertram F. Malle. 2004. *How the Mind Explains Behavior: Folk Explanations, Meaning, and Social Interaction.* The MIT Press, Cambridge, MA, USA.

[25] Bertram F. Malle and Joshua Knobe. 2001. The distinction between desire and intention: A folk-conceptual analysis. In *Intentions and Intentionality: Foundations of Social Cognition*, Bertram F. Malle, Louis J. Moses, and Dare A. Baldwin (Eds.). The MIT Press, Cambridge, MA, USA, 45–67.

[26] Abraham Harold Maslow. 1943. A theoory of human motivation. *Psychological Review* 50 (1943), 370–396.

[27] Hartmut Rosa. 2019. *Resonance. A Sociology of Our Relationship to the World.* Polity, Cambridge, UK.

[28] Raymond Ka-Man Sheh. 2017. "Why did you do that?" Explainable intelligent robots. In *Proceedings of the Workshops of the 31st AAAI Conference on Artificial Intelligence*. San Francisco, CA, USA, 628–634.

[29] Ron Sun. 2009. Motivational representations within a computational cognitive architecture. *Cognitive Computation* 1 (2009), 91–103. https://doi.org/10.1007/s12559-009-9005-z

[30] Toby Tyrrell. 1993. *Computational mechanisms for action selection.* Ph.D. Dissertation. University of Edinburgh, Edinburgh, UK. https://doi.org/1842/20257

[31] Eva Wiese, Giorgio Metta, and Agnieszka Wykowska. 2017. Robots as intentional agents: Using neuroscientific methods to make robots appear more social. *Frontiers in Psychology* 8 (2017), 1663. https://doi.org/10.3389/fpsyg.2017.01663