KI - Künstliche Intelligenz (2022) 36:237–254
https://doi.org/10.1007/s13218-022-00772-8

TECHNICAL CONTRIBUTION

What's on Your Mind, NICO?

XHRI: A Framework for eXplainable Human-Robot Interaction

Matthias Kerzel[1] · Jakob Ambsdorf[1] · Dennis Becker[1] · Wenhao Lu[1] · Erik Strahl[1] · Josua Spisak[1] · Connor Gäde[1] · Tom Weber[1] · Stefan Wermter[1]

Received: 11 March 2022 / Accepted: 28 June 2022 / Published online: 17 August 2022

Abstract
Explainable AI has become an important field of research on neural machine learning models. However, most existing methods are designed as tools that provide expert users with additional insights into their models. In contrast, in human-robot interaction scenarios, non-expert users are frequently confronted with complex, embodied AI systems whose inner workings are unknown. Therefore, eXplainable Human-Robot Interaction (XHRI) should leverage the user's intuitive ability to collaborate and to use efficient communication. Using NICO, the Neuro-Inspired COmpanion, as a use-case study, we propose an XHRI framework and show how different types of explanations enhance the interaction experience. These explanations range from (a) non-verbal cues for simple and intuitive feedback of inner states via (b) comprehensive verbal explanations of the robot's intentions, knowledge and reasoning to (c) multimodal explanations using visualizations, speech and text. We revisit past HRI-related studies conducted with NICO and analyze them with the proposed framework. Furthermore, we present two novel XHRI approaches to extract suitable verbal and multimodal explanations from neural network modules in an HRI scenario.

Keywords Explainable AI (XAI) · Neuro-robotics · Human–robot interaction · Trust in artificial intelligence

1 Introduction

According to the concept of Theory of Mind (ToM) [43], humans have the intuitive ability to understand the intentions, knowledge, and reasoning of an interaction or collaboration partner by ascribing mental states and thoughts to their opposites. For these assumptions to arise, the other person does not need to explicitly communicate their thought or action. Instead, humans rely on interpreting non-verbal cues such as gestures and facial expressions, as well as high-level verbal explanations of intentions, actions, and prior assumptions, before requiring more detailed explanations of a reasoning process. It is likely that such mechanisms also play a significant role when humans collaborate with robots. Therefore, we argue that explanations in Human–Robot Interaction (HRI) should encompass a range of communication and interpretation methods, from the ability to express non-verbal cues to formulating possible issues on a high level, and explaining underlying AI processes in-depth. To this end, we propose a framework for eXplainable Human–Robot Interaction (XHRI).

The XHRI framework covers three stages of explanations: (1) intuitive and fast communication with non-verbal cues like gestures, gaze, facial expressions, and non-speech sounds to establish rapport and signal attention or possible issues, (2) linguistic high-level explanations of the robot's intentions, actions, assumptions as part of its symbolic component, (3) detailed, in-depth explanations of neural and machine learning modules used by the robot that rely on non-symbolic mediums such as visualizations. As an example, consider the scenario shown in Fig. 1: A user asks the humanoid robot NICO to hand over the lemon. Initially, NICO confirms that it understood the task and will carry it out. Subsequently, NICO uses non-verbal communication by displaying a stylized surprised facial expression to indicate an issue. The person recognizes immediately that a problem

✉ Matthias Kerzel
matthias.kerzel@uni-hamburg.de

[1] Knowledge Technology Group, Universität Hamburg, MIN-Fakultät, Fachbereich Informatik, Vogt-Kölln-Straße 30, 22527 Hamburg, Germany

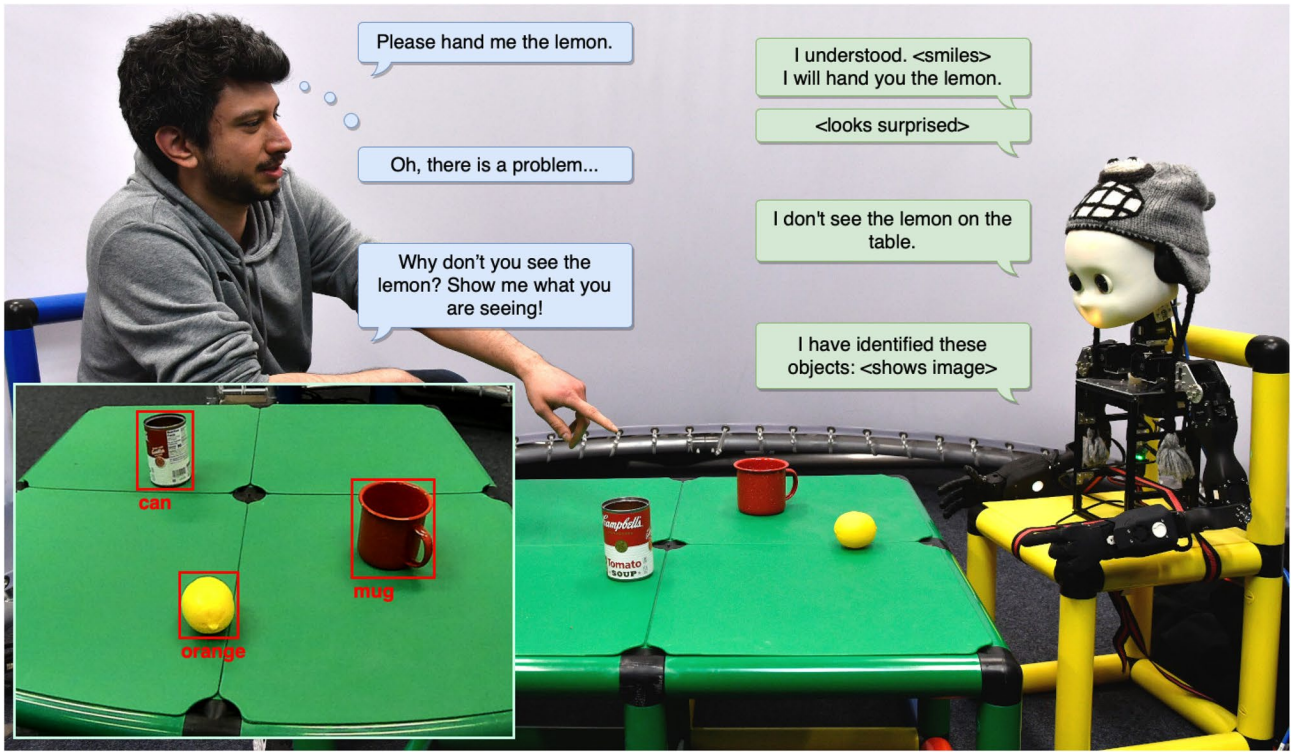

Fig. 1 NICO, the humanoid robot, assists a person. NICO uses non-verbal communication to signal to have an issue with a received request; it can explain the issue on a symbolic level and, when prompted, can also visualize elements of the underlying neural processing. The visualization is accompanied by a verbal explanation of what is shown and requires an external display, disrupting the human–robot interaction. Therefore, visualizations should only be used when necessary and complemented by a verbal description of what is shown. The NICO robot used in this task consists only of the upper body, see Fig. 3 for the fully-equipped NICO

occurred, and NICO explains, on a high level, that it cannot identify the lemon. Since the person is positive that there is a lemon on the table, he asks NICO to explain what objects the robot is recognizing on the table.

We apply this framework to revisit a series of past HRI-related studies that were conducted on the child-sized humanoid robot NICO [31], the Neuro-Inspired COmpanion. While these studies were originally aimed at diverse HRI-aspects, we re-analyze them in this article from the perspective of explainability. We show that both non-verbal communication and high-level linguistic explanations affect the human-robot interaction positively. We also analyze a wizard-of-oz-like study that shows the positive effect of providing in-depth explanations in natural language for a neural reasoning component.

Based on these findings, we present two specific novel approaches for XHRI: First, we extend the neural object detector RetinaNet [41] to realize an explainable object-picking scenario, similarly as shown in Fig. 1. By utilizing the detector's confidence and a novel reasoning module, we create a hybrid neuro-symbolic object detector. Second, we apply a neural explanation method to an end-to-end grasp-learning approach to analyze which part of the visual input is relevant for the robot's motor actions. These two XAI-abilities can enable robots to explain their actions and decisions to human interaction partners in many contexts.

In summary, our contributions are as follows:

- A hybrid neuro-symbolic framework for XHRI based on Theory of Mind.
- A re-analysis of past HRI studies on the NICO robotic platform. The results are interpreted within the XHRI framework, and conclusions are drawn for the effects and design of non-verbal, high-level, and in-depth explanations.
- Two novel approaches for XHRI that extend a neural object detector and a neural end-to-end grasp-learning approach with the ability to explain their processing.

The remainder of this article is structured as follows: In related work, a summary of Theory of Mind and an overview of XAI methods that can be applied to neural approaches is provided. The methodology section introduces the proposed XHRI framework, which is used to structure the review of existing HRI studies and novel XAI methods. Additionally, a neuro-symbolic control architecture and the humanoid

robot NICO are presented. The NICO robot was employed in various HRI studies and is used to introduce two novel XHRI approaches. Finally, we summarize the results, contributions, and future work in the conclusion.

2 Related Work

In the following, Theory of Mind is introduced as a foundation for our eXplainable Human-Robot Interaction (XHRI) framework. Then, a brief overview of eXplainable AI in HRI is presented, before introducing XAI methodologies for explaining neural networks.

2.1 Theory of Mind in HRI

A large body of research in HRI and psychology established evidence for humans assigning anthropomorphic attributes to inanimate objects such as robots and artificial agents [20, 21, 35]. Notably, humans seem to attribute thought processes to agents, as formulated in the conceptual framework of "Theory of Mind" (ToM) [43]. This enables users to predict the robots' behavior and model their mental states, such as beliefs, desires, intents and emotions, similar to how they perceive a human interaction partner. A review of human–robot versus human–human interaction is given by Krämer et al. [35], while Hellstrom and Bensch [21] provide an overview of ToM in robotics with respect to "understandable robots".

The assumption that humans employ ToM during interactions with robots is supported by preliminary results from functional neuroimaging studies [19]. Banks [4] replicated an established test-suite for assessing implicit signals for Theory of Mind towards humans with robots, showing that participants exhibit similar implicit signals as long as social cues expressed by the robots are sufficiently similar to those used by humans and interpretable.

Empirical studies with artificial agents and robots are in line with this view, suggesting that humans exhibit behaviour in human-robot interaction that is similar to human-human interaction [35]. For example, the authors of [38] demonstrate that humans build a mental model of a robot's factual knowledge, extrapolating from their own knowledge, and taking into account information about the origin of the robot. A study on a robot's reaction to being touched by participants [49] show that a passive movement after touching can evoke the impression of familiarity towards the participant, while a repelling movement induces the attribution of intentionality, and no reaction led to a negative impression of the robot. In a qualitative evaluation of a museum agent [34], users employed human-like conversation methods, with the majority of users greeting the system and about a third saying good-bye, despite the fact that they could just walk away from the agent. Furthermore, many participants of that study asked anthropomorphic questions that seemed to assume or probe human-like beliefs, emotions and thoughts, such as "How are you?", or even "Are you in love?". Other studies reported a tendency of users to communicate freely with agents, even if they were told that they would understand only a strictly predefined set of verbal commands [24, 54].

While we focus on human Theory of Mind towards robots in this work, there has been longstanding research arguing for the need to implement a ToM towards humans in robots as well, to enable the prediction of human behavior and act depending on the user's context [11, 56]. In the context of explanations, it has been argued that modelling the user's mind properly is an integral part of a successful explanation and that considering the perspective of the explainee can influence human-human explanations considerably [45].

2.2 Use of Explainable AI in HRI

Through the widespread success of deep learning in multiple disciplines, the technology quickly found its way into robotics. Deep neural approaches significantly increased performance in perceptual problems relevant to the field, such as computer vision [65] or natural language processing [68], but also control and planning [57]. However, a major criticism of deep neural networks is the lack of transparency, as their decision-making process is inherently difficult to explain. In an attempt to derive interpretable models, various methods emerged under the term *eXplainable AI* (XAI) [3, 15]. While these approaches succeeded in making models more transparent to some extent, most of the developed approaches fail to offer explanations that are understandable for a lay user [46]. Yet, in the field of human-robot interaction [60], the need for explaining and communicating a robot's actions and intent in an intuitively understandable manner has been pronounced by many authors [2]. Furthermore, XAI approaches outside HRI are typically concerned with data-driven models [3] rather than autonomous, goal-driven agents. A review by Anjomshoae et al. [2] on explainable agents and robots demonstrates that most work in this area either presents conceptual studies, lacks evaluation, or focuses on simplistic scenarios only, highlighting the lack of holistic and empirically validated approaches in the field.

This stands in contrast to longer-established research in HRI on how to effectively communicate information to increase the understandability of robots, using different terms such as readability [63], legibility [40], transparency [67], predictability [10], and many others [21, 55]. The main benefits of explainable HRI mentioned in the literature are increased trust and confidence in the system, as well as higher performance and efficiency in collaborative tasks. Additional drivers include teaching explanations for educating users about the robot and debugging in case of errors

and failures [2, 59]. Contrary to developments in XAI and deep learning, Sheh [59] found that existing work in effective communication of explanations in HRI predominantly assumed implementations based on white-box models, therefore disregarding any of the recently developed explanation methods for deep neural models. We address this gap in the state of the art with the proposed framework for eXplainable Human-Robot Interaction and two examples for explaining black-box neural approaches.

2.3 Explaining Neural AI

Explaining Data-driven Neural Learning. The recent development of XAI in deep learning (i.e. mainly supervised learning) focuses on enabling an understanding of a trained network; thus, many XAI approaches fall into the category of local, post-hoc (passive) attribution explanation [53, 58]. For example, in computer vision tasks, the explanation is usually represented as a saliency map [62]. The main principle of saliency maps is to identify parts of the input that are most influential for the decision derived by the neural network. Among the pioneer works in XAI is Class Activation Map (CAM), which was proposed by Zhou et al. [70] to highlight class-specific discriminative regions in the input image.

However, a drawback of this method is that the network's architecture has to be modified and the network retrained. This shortcoming is addressed in Grad-CAM [58] which instead leverages the gradient information of the neural network to identify relevant image regions for the obtained predictions. Along the direction of the gradient, the algorithm allows highlighting input-specific sub-areas that contribute to the predicted class label. A similar approach is LIME [53], which is a model-agnostic explanation approach that does not require adjustment of the model architecture.

These visual explanation techniques can be applied to most existing vision tasks and can provide valuable benefits to human-robot interaction [12]. Specifically, when the participant is curious or reliant on the recommendation or action of a robotic partner, an explanation can provide reassurance to the participant.

Explaining Reinforcement Learning. While most lines of XAI works in the field of supervised learning are applicable to image classification [33, 61] and visual question answering [58], explaining long-term decisions in reinforcement learning (RL) [64] remains challenging. Specifically, the additional temporal dimension of the episodic learning in RL hinders the use of explainable methods. However, recent success in reinforcement learning is a result of prior works conducted in deep learning [37]. Similarly, saliency maps to explain the decision-making in Deep Reinforcement Learning (DRL) have been adopted and promise to provide valuable insights into a model's reasoning process [17, 22, 69].

However, a common drawback of such methods is that only the agent's local input (state) is considered for decision-making. These methods neglect that perturbation-based explanations can introduce noise in the distorted regions that are irrelevant for the action taken by the agent [52]. Also, these visualization techniques are not necessarily understandable to a layman [47].

A different direction of explainable RL utilizes reward decomposition as a metric to derive an explanation of the robot's actions [25]. An agent's reward is factored into a sum of meaningful reward types which can be associated with the selected action. Although this allows to explain why one action is preferred to another, this approach of explaining RL is currently limited to Q learning [64], which relies on a discrete action space and deterministic dynamics.

To increase the faithfulness of explanations for RL, the causality of actions is considered [26, 42]. These methods enable learning a causal graphical model [14] while learning an optimal policy. In contrast to the attribution explanation mentioned, the causal graphic model seeks to reveal the model's functions and behaviour. In practice, this causality property is of great interest to HRI since a practitioner might be more concerned with an illustrative graph with a contrastive explanation (e.g., What-if question). This would allow the robot to respond to why one action was preferred over another. This strengthens our argument that XRL methods in HRI require trustworthy and accountable systems, which could shift the field's attention towards the development of interactive methods and post-hoc RL explanations.

3 Methodology

First, we propose the novel eXplainable Human-Robot Interaction framework XHRI that allows to categorize types of explanations in human-robot interaction. Then, the structure of hybrid neuro-symbolic control architectures that are promising for social robots is outlined. Finally, we introduce the Neuro-Inspired COmpanion (NICO) [31] which is utilized in various neuro-robotic and HRI experiments.

3.1 Framework for eXplainable Human-Robot Interaction (XHRI)

A robot that is interacting and collaborating with humans in a domestic or other dynamic environment has to explain its behaviour. The explanation aims to establish trust and enable a smooth collaboration by informing the user about the robot's state, beliefs, desires, and intentions. However, explanations can be obtrusive, time-consuming, and interfere with the robot's actual task. Therefore, like in a human

Table 1 Explanation techniques for symbolic and neural modules across different modalities in the XHRI framework. Approaches for explaining neural modules, particularly in the robotics domain, are emerging research. Two novel multimodal explanation techniques for explaining neural models in HRI are presented in Sect. 5

	Non-verbal	Verbal	Multimodal
Symbolic	Abstract symbolic information expressed e.g. as gestures, emotions, gaze	Language generated from symbolic information	Data visualization (e.g. graphs) with verbal instructions and context description
Neural	Learned emotion expression, extracted symbolic information (e.g. surprisal, error detection)	Rationalization generation using language models, extracted symbolic information (e.g. detected objects, confidence)	Neural XAI techniques (e.g. saliency maps) with verbal instructions and context description

collaboration, an efficient communication of the explanations is required.

Humans ascribe mental states, thoughts, and intentions to other people, forming a Theory of Mind, and employ this concept similarly when interacting with robots (see Sect. 2.2). From the point of view of HRI, this finding is promising insofar as robots are perceived naturally and intuitively, analogously to a human interaction or collaboration partner. However, on the other hand, the projection of thoughts and intents on the robot implies the possibility of the attribution of false assumptions on the side of the human. Therefore, we argue that the two fundamental functions of explanations in social HRI are to (a) communicate the model of the robot's current state to supply the interaction partner with information to construct a meaningful ToM, thus enabling an intuitive perception of the robot, and (b) update and correct the humans' assumptions continuously. If both aspects are sufficiently fulfilled, a natural, predictable, trustful, and efficient human–robot interaction can be achieved.

However, explaining a robot's behavior can be challenging. Typically, modern robots are realized using a neuro-robotic architecture that consists of different symbolic and sub-symbolic (neural) components, rather than being a monolithic system. When explaining the state of the robot to users, it requires not only different explanation techniques for the distinct systems, but also different communication modalities, as illustrated in Fig. 2. In the following, we distinguish between three levels of communication: non-verbal, verbal, and multimodal explanations. A summary of these categories is provided in Table 1.

Non-verbal Explanations can be a powerful feedback mechanism, especially for interactions with non-expert users and in noisy environments. Gestures, emotions, or gaze provide intuitive and non-obstructive feedback during the interaction. This can also appear in parallel to ongoing verbal communication. Using non-verbal communication, the user can be quickly informed about the robot's responsiveness, even before slower, in-depth processing is completed. For example, the robot can look at a person that is speaking, signaling its attention, even before processing the contents of the speech. Information from symbolic and neural subsystems can be abstracted and communicated on this level. Additionally, a robot should be able to use this communication in both directions: it needs to be able to produce non-verbal cues and to react to them. However, the expressiveness of non-verbal explanations is limited and not suited to communicate complex reasoning processes.

Verbal Explanations can be used on a symbolic level to summarize and verbalize the robot's current state, announce its next action, or communicate its knowledge and observations. While some symbolic agents have a considerable complexity and would require equally complex explanations, the main focus at this level is to answer and explain the robot's actions, intentions, and knowledge on a high level in an accessible manner. Ideally, complex relationships in symbolic systems can be explained faithfully in natural language. For example, a robot can inform the user about its assumptions, as queried from a knowledge base, its logical reasoning steps, and derived conclusions. Emerging research in XAI focuses on extracting symbolic

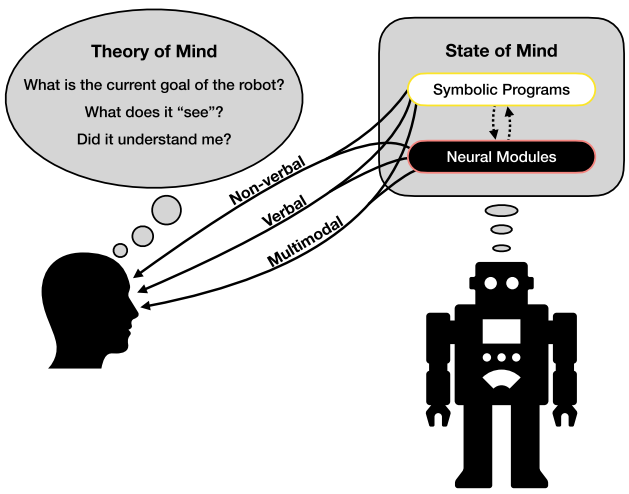

Fig. 2 The relationship between the user's Theory of Mind, the robot's state, and means of communicating information from different components in a neural-symbolic robot architecture

explanations from neural networks and communicating them in natural language. Due to the understandability and expressiveness of these explanations, we argue that developing these approaches further and deploying them to social robots will be crucial for achieving explainable HRI.

Multimodal Explanations are often required to convey explanations for neural modules and rely on the XAI methodology as described in Sect. 2.3. On this level of explanations, a robot should be able to explain the inner workings of its neural modules and why a certain input has led to its current actions, outputs, or decisions. Such neural models, with millions of learnable parameters, cannot be explained directly in terms of their computations. Adequate visualization- and explainability techniques have to be developed that can seamlessly be integrated with the previous two levels of explanations. As current approaches for neural network explanation largely rely on visualization techniques [62], they require additional display outputs. Further, inferring information from visualizations requires substantial interpretation on the side of the user, opening up the possibility of misunderstandings, especially with non-expert users. We argue that due to these limitations, visualizations should not be employed by themselves when interacting with non-expert users, but rather in conjunction with information assisting in interpreting the visualization in the form of a multimodal explanation. In the case of symbolic subsystems, visualizations can be leveraged to summarize information, for example, in diagrams or graphs, together with a verbal explanation.

In summary, explainable AI in the context of human-robot interaction can be communicated on three levels:

- Intuitive and non-verbal feedback and communication in both directions between the robot and the human interaction partner.
- Verbal explanations in natural language.
- Multimodal explanations using visualizations accompanied by verbalization and assisting information.

We argue that, when communicating explanations, it should be done in the most intuitive, unobtrusive, and efficient way possible before escalating to more complex explanations. As illustrated in Fig. 1, this can start with abstract, non-verbal communication to signal an issue, continue with a verbal statement providing a high-level explanation, and end in a detailed multimodal explanation with a visualization to explain the inner workings of a neural module. Such a visualisation of the neural module should be accompanied by verbal explanations.

3.2 A Neuro-robotic Architecture

Neuro-robotic architectures are hybrid neuro-symbolic systems. The architecture depicted in Fig. 4 is representative of the neuro-robotic architectures realized on robots such as the NICO (Sect. 3.3), which was employed in various HRI studies. A central symbolic agent controls the main behavior of the robot and the execution of neural modules. This main symbolic agent consists of a reasoning mechanism and a knowledge base; it can be realized as a state machine, a logic program, or a sequence of imperative commands. This central component enables complex robot behavior. Even though most of a robot's functionality can be realized with neural modules, the symbolic component manages a variety of processes and data distribution, e.g., recording images from the cameras, preprocessing them, and forwarding the image to a neural network module. Furthermore, the symbolic agent manages the loading process and distribution of computational results or neural modules. This symbolic component is a *white box* in the sense that its state, processing, and knowledge can be queried.

In contrast, the neural modules act as *black boxes*, their processing is intransparent due to their complexity. Neural modules often form a bridge between the robot and the physical environment by realizing sensing or actuation functionality. Input from a robot's sensors is processed by the neural modules, and the derived information can be integrated as symbolic knowledge into the symbolic agent. Likewise, symbolic knowledge can be transformed by neural modules to create motor or actuator commands.

Examples of obtaining symbolic representations from sensory input are identifying and locating objects in an image, recognizing the facial expression of a human face, generating text from an audio signal, or identifying objects by their characteristic sounds or haptic properties. Similarly, an action derived from a symbolic representation can be translated into a motor command to grasp an object, a facial emotion that is displayed via LEDs, or speech that is generated from text. Neural modules can also perform reasoning tasks without directly interacting with the robot's sensors or actuators. For example, models that learn to play a game and decide which next move is ideal [1]. Finally, neural end-to-end models directly transform raw sensory information into actions, such as end-to-end grasp learning [32]. However, even for these modules, there is usually a symbolic agent controlling the overall process.

This neuro-robotic control architecture is utilized in the NICO robot and is reflected in our proposed framework for eXplainable Human-Robot Interaction.

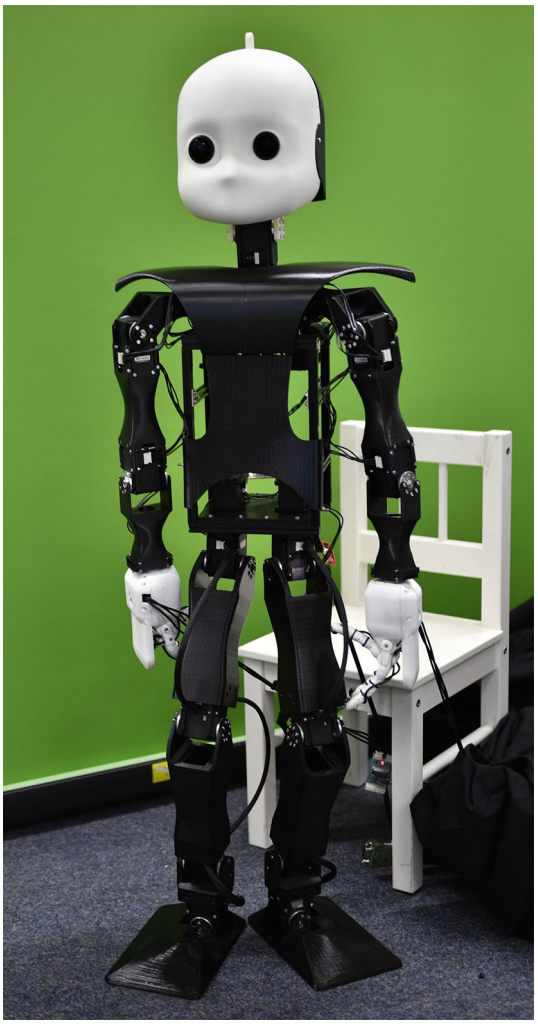

Fig. 3 NICO, the Neuro-Inspired COmpanion

3.3 NICO, the Neuro-Inspired COmpanion

NICO [31] (see Fig. 3) is designed to represent a child-sized humanoid robot that enables research on embodied neuro-cognitive models. The robotic design and software components are open-source[1]. The design of this robot combines aspects of humanoid and social robotics. The robot's sensors comprise vision, audio, and haptic perception. NICO's arms and hands allow human-like movements and object manipulation. In conjunction with the overall human-like proportions, the movable head, hands, arms and option to display stylized emotions make NICO well-suited for HRI studies.

NICO's appearance and proportions roughly follow that of a young child with a height of about one meter. NICO's head is an adaptation of the open iCub design, featuring a stylized human appearance. The frame-like body of NICO can either be covered in clothing or with 3D-printed solid parts, as shown in Fig. 1.

NICO's motor abilities mimic that of a human with 30 DoF that control head, arm and leg motion: To perform gaze shifts and to signal attention, NICO's head can perform yaw and pitch motions. NICO's arms have six DoF, three in the shoulder, one in the elbow, and two in the wrist to perform gestures and to grasp and manipulate objects. The arms end in Seed Robotics' RH4D-articulated hands with three fingers, which can be exchanged for compatible end-effectors. Finally, NICO has legs with six DoF each.

NICO's sensing abilities encompass sight, hearing and haptic perception. Two See3CAM CU135 cameras with a wide-angle and a 4K resolution are installed in NICO's head. For sound perception, NICO can be equipped with pinnae which embed Soundman OKM II binaural microphones. Haptic sensing is realized with a combination of feedback from the robot's motors to register the current load on each DoF and OPTOFORCE OMD-10-SE-10N7 force sensors at the tips of the fingers.

Non-verbal communication can be realized not only with head motion and arm and hand gestures; NICO can also display stylized emotions and other patterns via LED arrays in the mouth and eye region. Additionally, an integrated speaker allows verbal communication.

Overall, NICO is designed to be an accessible, affordable, open-access platform for supporting research on explainable human-robot interaction.

4 Re-analysis of HRI Experiments with NICO from an XAI perspective

In this section, we will re-analyze and summarize a series of past HRI experiments conducted with NICO from the perspective of our proposed XHRI framework. While most of the selected experiments are not originally focused on XAI, they either highlight examples for the implementation of individual elements that are required in the XHRI framework, or demonstrate effects of explanations and communication modalities. The studies are re-interpreted in the context of XAI. For ease of reading, we provide a summary of the analyzed studies in Table 2.

The studies reported below by Kerzel et al. [28] and Churamani et al. [9] are concerned with displaying and learning emotions, an important prerequisite for communicating non-verbal explanations when viewed through the lens of the XHRI framework. In [71], the effects of non-verbal communication on trust, a main driver for explainable AI research in HRI, are examined. The work summarized in [6] combine verbal utterances and non-verbal cues to

[1] Further information, open-source CAD files, and the NICO API are available at https://www.inf.uni-hamburg.de/en/inst/ab/wtm/research/neurobotics/nico.html.

communicate emotions, demonstrating the utility of multimodal communication. As an experiment in verbal communication, the study [27] demonstrates the effect of explaining the current state of the robot in a grasping task by verbalizing symbolic information derived from a state machine controlling the robot. In [8], users are involved in a personalized dialogue, for which symbolic information is extracted from neural modules and stored in a knowledge base, illustrating an approach for lifting sub-symbolic information to the symbolic space, where information can be expressed verbally. Finally, the effects of verbal explanations on how a robot is perceived are explicitly assessed in the study [1], where the strategy behind the current move in a competitive board game is explained. For details on the experiments, we refer to the original articles; in this paper, we highlight those aspects of the studies relevant for XHRI.

Furthermore, the studies reported below illustrate different ways to communicate the internal state of a robot and its intentions, plans and requests to users. In summary, the studies highlight that different communication channels are necessary for an intuitive HRI-experience and the communication of explanations. However, the studies also highlight a need for computational XAI approaches for extracting explanations from sub-symbolic modules of a robot, e.g., neural networks, that are suitable for non-verbal and verbal explanations.

4.1 Non-verbal Communication: The Role of Displaying and Recognizing Emotions in XAI

We summarize four studies that demonstrate the effect of NICO recognizing or showing emotions in an HRI scenario.

Showing Emotion Expressions: Facial expressions are an essential aspect of non-verbal communication. Besides universal communication of information, they can be used in parallel to verbal communication. NICO can display a variety of emotions with the LEDs located at the mouth and eyebrow area. The recognizability of seven different emotions was evaluated with human participants by Kerzel et al. [31]. Twenty participants were asked to name the emotion that the robot displayed. A subset of 5 emotions (neutral, happiness, sadness, surprise, and anger) could be identified reliably by the participants.

For evaluating the change in perception of a robot that displays emotions with facial gestures, the participant's perception of NICO was assessed before and after interacting with the robot. After observing the displayed emotions of the robot, a significant increase in anthropomorphism, animacy, likeability, and safety was observed. The study shows that NICO is able to express basic emotions with the facial display, an essential prerequisite for non-verbal social communication within our XHRI framework. Furthermore, expressing this ability has an overall positive effect on how participants perceive the human-robot interaction.

Recognizing and Mirroring Emotion Expressions: In a study conducted by Churamani et al. [9], NICO was used not only to display emotions but also to recognize them. A convolutional neural network in conjunction with a self-organizing map was employed for perceiving the five emotions *anger*, *happiness*, *neutral*, *surprise* and *sadness*. Additionally, two Multi-Layer-Perceptron (MLP) networks were implemented for learning to mirror the emotions on NICO's face-LED display, with one network learning to recognize and mirror emotions in general while the other network was learning to display the emotions as expressed by a specific individual. Depending on the experimental conditions, NICO was able to recognize and mirror two to four of the five above-mentioned emotions, depending on how much the networks were trained with individual interaction partners.

The results demonstrate the potential to adaptively learn and express emotions in humanoid robots by learning from online interaction. A solution to not only continually learn [51] and improve this perception and communication of emotional states, but also to accommodate and adapt to inter-individual differences was implemented. Two important parts of human-robot interaction in relation to robot Theory of Mind, relevant to the XHRI framework, were addressed: First, learning to identify emotions of humans as a basis for forming a model of the condition of the human mind, and second, communicating emotions to the human via facial expression.

Non-verbal Communication in a Cooperative Scenario: Zörner et al. [71] analyzed the influence of non-verbal communication in developing trust with a robot in a cooperative scenario. In the fictional experimental scenario, two NICO robots offer information and advice to the user who has to assign (fictional) limited resources to the different actions suggested by the two robots as a measurement of trust. After the experiment, the Godspeed questionnaire and Risk Propensity Scale [44] are assessed.

For the evaluation of the experiment, the data of 45 participants were analyzed. The Godspeed questionnaire reveals a significantly greater rating in anthropomorphism and animacy of the robot that communicated with non-verbal cues. This study further confirms the importance of non-verbal communication and that it can foster trust in human-robot interaction. From the point of view of XHRI, this finding is especially relevant, as trust is one of the main reasons for the development of explainable robots. Therefore, for the goal of fostering trust in robots, non-verbal communication requires careful consideration alongside explanation contents.

Emotionally Engaging HRI: To assess the change in human perception of a robot expressing emotions, the difference between a socially engaging and a competitive robot was researched by Beik-Mohammadi et al. [6]. In the

experiment, a human and NICO played a competitive game. In the between-subject study, the socially engaging robot utilizes NICO's facial LEDs to express emotion, gestures, and different voice pitches to convey emotion. The competitive robot did not utilize these social features and was portrayed as rational.

During the experiment, the sentiment of the participants was automatically analyzed with face recognition and a neural network that estimates the sentiment of the participant's face. After the experiment, the Godspeed and Mind Perception [16] questionnaires were used to assess the differences in perception of both settings. Twenty-two participants took part in the study.

The evaluation of the Godspeed questionnaire shows that the socially engaging NICO was perceived as more likable and received higher animacy ratings than the competitive robot. The analysis of the expressed emotions during the interaction with the robot revealed that the participants displayed a wider range of emotions when interacting with the socially engaging robot. Specifically, a significant increase in happiness was noticeable, in contrast to the interaction with the competitive robot, which was dominated by a display of neutral emotions.

Overall, the robot displaying emotions was more engaging for the participants. From an XAI perspective, the results, again, suggest that using non-verbal communication strongly influences the perception and willingness to interact with a robot, which is vital for successful explanations in an HRI context.

4.2 Verbally Explaining Symbolic, Hybrid Neurosymbolic and Neural Reasoning Processes

After having reviewed studies that use non-verbal explanations, we present a summary of three studies that (mainly) use verbal communication to explain the symbolic, neuro-symbolic or neural AI behind the robot's actions.

Active Requests and Explanations in Grasp Learning—Symbolic Reasoning. Kerzel et al. [29] conducted an HRI study where non-expert participants assisted NICO in learning how to grasp, either instructed by a human experimenter or by NICO. The participants helped to collect data samples for a neural end-to-end architecture for grasp learning. In this approach, the visuomotor ability of the robot is learned in a supervised way, based on samples the robot collects semi-autonomously from interaction with the environment [32]. The approach is semi-autonomous, as the robot can, in principle, continuously collect samples on its own, but initially, an object must be placed into NICO's hand and occasionally, objects might slip from the robot's hands or roll off the table. This problem could be detected by the robot using haptic sensing, but human assistance is required to place the object back into NICO's hand.

A state machine controls the image recording, motor actions, and haptic sensing to ensure that a sample is collected correctly. Additionally, the state machine uses the facial emotion display and triggers speech output to keep the interaction partner informed about its current state and to request assistance. During the experiment, NICO explains to the participants its actions, *"I will try to grasp the object. Oh no! I failed to grasp the object."*, request assistance, *"It's training time. Place the training object in my right hand."*.

It could be shown that in the condition where NICO communicated directly with the participants, the users were much more engaged in the learning process. This was not only reflected in their subjective assessment of NICO as having a higher perceived safety, animacy, and anthropomorphism, but also in the behavioral measure of how often the participants physically interacted with the robot or the grasp object on their own initiative to ensure NICO can handle the object without error. Re-interpreted in an XAI context, the study provides evidence that a robot explaining its actions and goals has a positive effect on the user's trust and engagement. In this study, the explanations were extracted from the symbolic control realized as a state machine.

Personalized Dialogue Using Symbolic Knowledge Bases—Hybrid Neuro-symbolic Reasoning. NICO was utilized to study the effects of a personalized dialogue system, where the robot remembers previous users and personalizes the communication accordingly [8, 50]. In an object-learning scenario, users taught NICO to recognize the location of objects on a table. 27 participants were divided into two groups. While the first group was presented with a baseline scenario of only the teaching task, the second group was engaged in a personalized interaction with NICO prior to the task. In this condition, the robot asked for personal information such as their name and nationality while tracking and learning their faces. The personal information was stored in a symbolic knowledge base. After the teaching task, in a subsequent second interaction, NICO recognized participants by their faces and way of speaking. NICO greeted them by their name and involved the users in another short personalized dialogue, retrieving the previously stored information.

The perception of the robot in both conditions was assessed using a Godspeed questionnaire [5] and a customized version of the UTAUT questionnaire [39], alongside questions regarding NICO's attentive and recognizing capabilities. On the Godspeed questionnaire, the group involved in the personalized dialogue gave significantly higher ratings in likeability and safety, while the UTAUT dimension of social acceptance was perceived as worse compared to the baseline condition.

This work is an example demonstration of how symbolic information can be extracted from neural models and stored

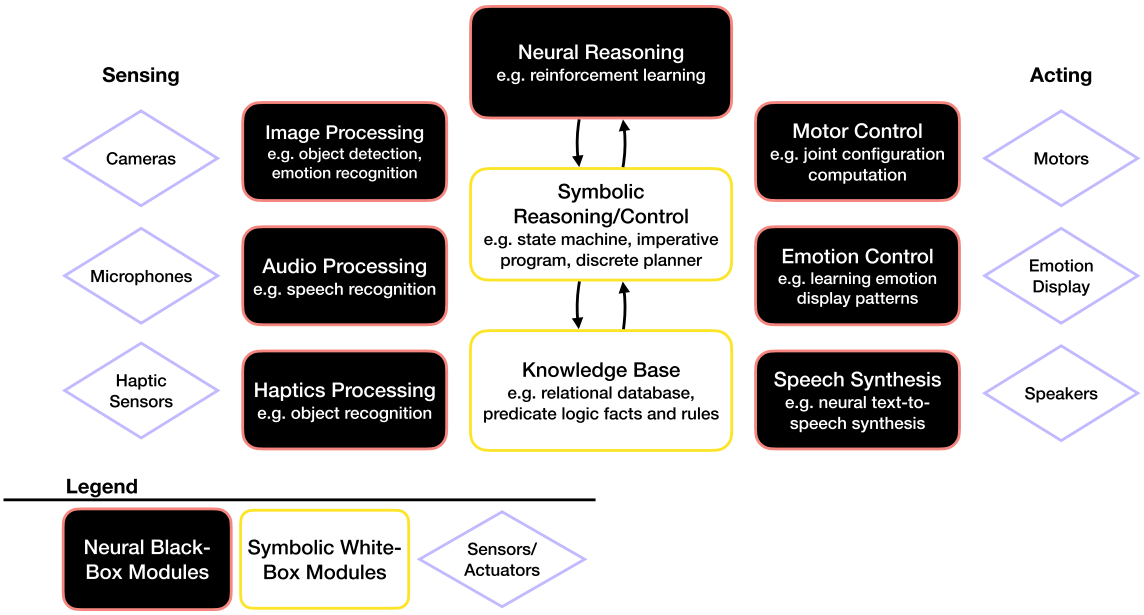

Fig. 4 Hybrid neuro-symbolic control architecture for robots. While the symbolic *white-box* modules are accessible for inspection, the neural *back-box* modules require specialized XAI approaches

in a symbolic knowledge base to be afterwards communicated in natural language, thus forming a hybrid neuro-symbolic system. While the study focused on enhancing the interaction through personalization, information lifted from neural models and stored in a symbolic knowledge base can likewise be used for generating verbal explanations. Additionally, in the XHRI context, personalizing dialogue is of particular importance for adapting explanations to a user, which is an important aspect of explanations identified in the social sciences [45].

Explaining Strategies in a Competitive Board Game—Neural Network-based Reasoning. In [1], two virtual NICO robots played a competitive board game against each other. Both robots' actions were based on a neural reinforcement learning approach. In an HRI study, a human participant is observing the match. While one robot provided explanations for the actions, the other robot only announced the moves without providing any reasoning of how these might impact the game or the strategic advantage. The explanations about the robots' actions were communicated in natural language and provided reasons for the move that are also understandable for participants that are not familiar with the game. The human-likeness and general perception of the robots, as well as their perceived competence [66] and trust in the robots' ability to be victorious, were measured using the Godspeed questionnaire [5] and enhanced by questions (Fig. 4).

Both robots used the same reinforcement learning agent based on the Deep Q-Network model [48]. It was trained by playing the game against different greedy strategies as well as self-play. As explanations in natural language are required, post-hoc visualization approaches such as a heatmap of Q-values are insufficient. Therefore, a rule-based rationalization approach was employed, with hard-coded explanations. While these explanations are not faithful to the models' reasoning, they generate plausible natural language utterances to examine the effect of explanations in this scenario. An illustration of the scenario with two NICO robots playing the game of Ultimate Tic-tac-toe is shown in Fig. 5.

In total, data of 92 participants were evaluated. A Wilcoxon test of the anthropomorphism for the measures of the XAI ($M = 12.42$, $SE = 0.53$) and non-XAI robot ($M = 10.97$, $SE = 0.56$) reveals a significant difference. Similarly, the Kruskal-Wallis test [36] shows a difference in the measures of animacy for the two robots ($p = 0.008$). In an XAI context, the study demonstrates that a robot that verbally explains its actions significantly influences the perception of the robot. We would like to highlight that in the conducted study, the robot does not utilize non-verbal communication. The study shows the effect that robots that provide verbal explanations are perceived favorably.

4.3 Discussion

The previous HRI studies conducted on the NICO robot show that the general use of non-verbal communication can positively influence the perception of the robot in terms of animacy, anthropomorphism, safety and likability. Further, we re-interpret the result of the studies as indicative that explanations of a robot's actions contribute to perceiving the robot as more human-like in three of the studies presented. However, it is formally not possible to disentangle these

Table 2 Meta-analysis of HRI studies on the NICO robot within the XHRI framework. The effects reported are the results of the respective study and not necessarily direct effects of explanations

Study	Summary	Expl. Levels	Participants	Abilities used	Inventory	Effect
Kerzel et al. (2017) [31]	Participants rate motionless NICO with and without facial emotion display.	Non-verbal	20	Facial display	Godspeed	Anthropomorphism+, animacy+
Churamani et al. (2017a) [9]	NICO is learning, recognizing and displaying emotions.	Non-verbal	10	Cameras, facial display	–	–
Zörner et al. (2021) [71]	In a cooperative scenario, the participant has to choose between two solutions presented by two different robots. One robot displays emotions using non-verbal communication, while the other does not.	Non-verbal	45	Facial display, gestures	Godspeed, Risk Propensity Scale	Anthropomorphism+,animacy+,correlation self-assessed risk-taking behavior and trust measured
Mohammadi et al. (2019) [6]	Participants play a competitive game against either an emotional or competitive NICO.	Non-verbal, Verbal	22	Facial display, gestures, voice	Godspeed, Mind Perception	Animacy+,likability+
Kerzel et al. (2020) [29]	Usersteach NICO how to grasp according to NICO's verbal guidance.	Non-verbal, Verbal	24	Cameras, facial display, voice, arm and hand	Godspeed	Safety+,animacy+,anthropomorphism+
Churamani et al. (2017b) [8], Ng et al. (2017) [50]	Users are recognized and involved in a personalized dialogue.	Verbal	27	Cameras, microphones, voice	Godspeed, UTAUT	Likeability+,safety+,social acceptance-
Ambsdorf et al. (2022) [1]	Participants observed gameplay of two NICO robots competing in the game Ultimate Tic-tac-toe and bet on the winner. While one robot explained its moves, the other just announced them.	Verbal	92	Cameras, voice	Godspeed, Perceived Competence	Anthropomorphism+, animacy+

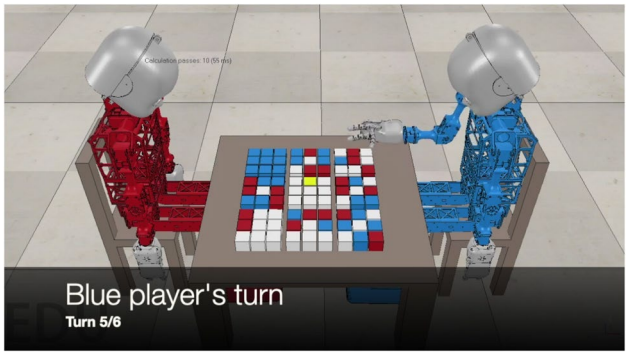

Fig. 5 Two humanoid robots play a game of UTTT against each other. Screenshot of a video sequence that was provided to the participants

effects post hoc. Studies examining the impact of integrating emotions and personality (i.e. Mohammadi et al. [6]) or dialogue personalization (Churamani et al. [8]) highlight the importance of incorporating insights from general human-robot interaction when designing explanations. In the experiment by Ambsdorf et al. [1], a wizard-of-oz-like approach for rule-based natural language rationale generation was implemented. Despite the lack of faithfulness in these explanations, they demonstrate a significant effect and highlight their importance in social robot interaction.

In summary, these results show the usefulness of multimodal explanations in HRI. However, they also show a need for XAI methods that can generate faithful natural language rationales that allow for communicating complex explanations in agreement with the model's decision-making or transform complex XAI visualizations for intuitive verbal and non-verbal communication.

However, state-of-the-art XAI approaches currently do not allow for meaningful information abstraction of neural models to express explanations verbally. Researchers and practitioners should consider multimodal explanations that combine visualizations with a verbal utterance or textual description. This could allow for more introspective explanations of neural modules and provide additional assistance to a user in interpreting the information presented. We suggest that extending these visualization methods by accompanying verbal and textual information is another promising research direction for HRI. To this end, we propose two new XAI methods in the next section.

5 Neural XAI Approaches for Human–Robot Interaction

As demonstrated in the previous section, re-analyzing HRI studies from an XAI perspective shows the importance of (a) non-verbal communication between a robot and a human collaboration partner, (b) explanations for the symbolic-level control of the robot and (c) explanations for the neural modules of the robot. However, the latter was only evaluated regarding its effect on the HRI without inspecting the actual underlying neural approach. In this sense, it was a wizard-of-oz-like study with hardcoded explanations instead of human-generated ones.

In this section, we will extend previous work with two novel approaches to incorporate neural XAI into a human–robot interaction scenario: the first approach explains a neural object detector by leveraging the detector's confidence and a reasoning mechanism to detect possible issues. The second example goes one step further and employs Grad-CAM to visualize and explain a neural end-to-end grasping approach.

5.1 Explaining Neural Object Detection in an HRI Scenario

Detecting objects, i.e. identifying objects and their location in the input image is a fundamental ability in robotics. For instance, Eppe et al. [13] use the neural object detector RetinaNet [41] to enable NICO to grasp objects on a table. RetinaNet classifies and localizes objects in an image in terms of an object class with classification confidence and a bounding box. However, this visual representation of the detected objects can become confusing due to overlapping detections and does not serve well as an explanation of why an object was or was not detected. To address this lack of explainability, we present an approach that leverages a combination of the detector's confidence and a reasoner to anticipate and explain potential issues in the robot's visual perception.

RetinaNet is a single-stage object detector, which uses a *focal loss* during training to balance a large number of negative samples against the small number of positive samples and to focus the network updates on challenging samples. RetinaNet's architecture consists of a convolutional ResNet [18] that is linked at different levels to a multi-scale Feature Pyramid Network (FPN) that has a high spatial resolution as well as complex visual features. From different levels of this FPN, subnetworks for object classification and bounding box regression are extended.

We use RetinaNet in a scenario where NICO picks objects from a table, as depicted in Fig. 1. For training, images with class and bounding box annotations are required. To avoid the need for human annotation, we employ a sim-to-real approach and train RetinaNet with 350 synthetic images of 18 different everyday objects from the YCB object set [7] created in Blender[2]. We train the image detector for 10

[2] https://www.blender.org/.

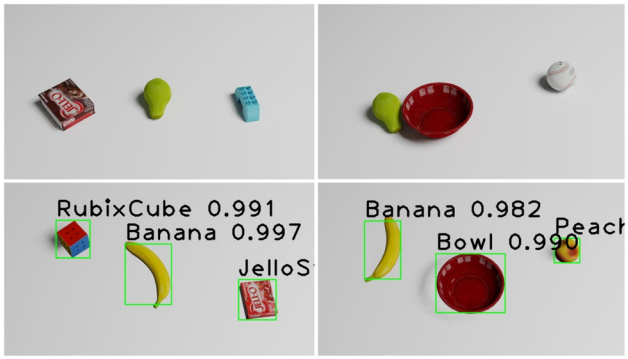

Fig. 6 Top row: Images from the synthetic training set. Bottom row: Images from the synthetic test set with bounding boxes and classification found by the trained RetinaNet

epochs based on a pre-trained ResNet50 using an Adam optimizer with a learning rate of 1e-5, the default batch size of 1, and a steps parameter of 350. We evaluate the model on 35 test images with a total of 105 object instances and achieve a precision of ~95% and a recall of ~100%, i.e. all objects are detected correctly, but a few objects are detected twice. As this result is sufficient, and we are mainly concerned with the XHRI aspects of this work, we avoid further optimization (Fig. 6).

We evaluate the approach on a set of 20 real-world images with 119 instances of everyday objects from the YCB object set. As expected, transferring a model from simulation to the real world causes issues: objects are detected multiple times, objects are detected but belong to the wrong class (false positives), and objects are not detected at all (false negatives). While the sim-to-real transfer could be improved using domain randomization, i.e. introduction of randomized variance with regard to colour, geometry, and other properties of the objects during training [23], we instead focus on improving the detector in an explainable manner.

Multiple and false detections can be addressed by filtering out detections with confidence below a fixed threshold; without such a measure, the number of false positives can be very large. Following the literature, we evaluate the approach for fixed thresholds of 0.5 and 0.7 and achieve a precision and recall of about 77% and 63% (0.5 threshold) and 92% and 59% (0.7 threshold). It can be observed that the higher the threshold, the more false positives are filtered out along with some true positives. The results also reflect the similar nature of some objects (i.e. two similar *Jell-O* packages; peach and lemon).

First, we address *multiple detections* of the same object in the image. If the intersection over union (IoU) of the bounding box of two objects is larger than 85%, only the detected object with the higher confidence is kept; the rest is discarded. This decision can be explained to the user, e.g., *"I found a lemon with confidence of 0.77, but it intersected with a peach which had a higher confidence of 0.93"*. Also, the information can be integrated into symbolic reasoning. In this example, if NICO is instructed to grasp a lemon but only recognizes a peach, it could further investigate. The results in Table 3 show that the recall and precision of this approach are slightly smaller; this is because double detections often contain the right classification along with an incorrect one. However, the table also shows that the total number of false-positive detections is lowered from about 13% to 11% for a threshold of 0.7 and from 26% to 21% for a threshold of 0.5, reducing the number of detected objects and the complexity of the explanation.

Table 3 Results for sim-to-real object detection with RetinaNet using thresholds of 0.5 and 0.7 without and with the IoU-based post-processing to avoid multiple detections of the same object

Setting	Precision	Recall	False Pos.
Threshold 0.5	~74%	~63%	~26%
Threshold 0.7	~88%	~59%	~13%
Thres. 0.5 IoU filter	~79%	~56%	~21%
Thres. 0.7 IoU filter	~89%	~53%	~11%

The second mechanism we introduce is a *double-threshold* for detection. Objects that are detected with a threshold of 0.7 are marked with a green bounding box, while those that are detected with a threshold of 0.5 are marked with a red bounding box to indicate a possible issue. This can also be verbalized to the user: *"Perhaps I found a red mug with low confidence of 0.56."*. Figure 7 shows an example output for image detection with both mechanisms and a sample verbalization.

We show that exploiting the confidence of RetinaNet and mechanisms along with a check for overlapping bounding boxes can generate verbal (i.e. text-only) as well as multimodal (i.e. text and image) explanations for a neural image detector.

Within our framework, this approach is an example for generating verbal explanations for a neural module from a robot's architecture. While the internal processing of the neural object detector is complex and can not be altered by a non-expert, the model's output and subsequent symbolic reasoning can be explained in plain terms; for instance, the choice of the object with the highest detection confidence in case of multiple recognitions or the rejection of a possible recognized object that a detection confidence below a threshold. Such explanations can help non-experts find solutions for perceptual issues by modifying the robot's environment, e.g., users could realize that particular objects are often confused and replace one of them with a more distinct object.

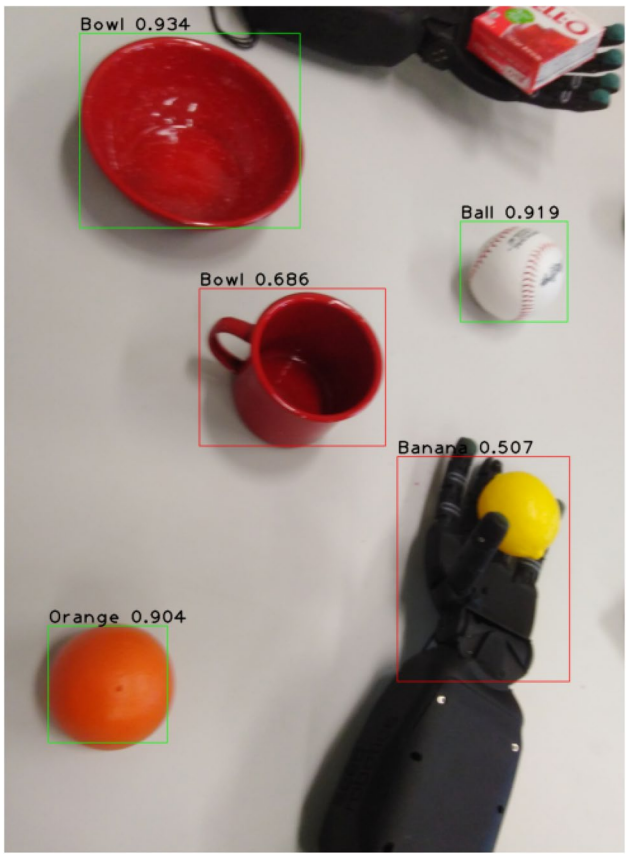

Fig. 7 Example of applying RetinaNet trained on simulated data to a real-world image. The system gives the following explanation (from top to bottom): *I found a bowl with high confidence of 0.93; I found a ball with high confidence of 0.92; perhaps I found a mug with low confidence of 0.68, but it was on top of a bowl which had slightly higher confidence of 0.69, so perhaps I found a bowl; I found an orange with high confidence of 0.9; Perhaps I found a banana with low confidence of 0.50*

5.2 Explaining Neural End-to-end Grasp Learning

Above, we utilize a neural object detector's confidence and an additional reasoning module to de-clutter and explain object detection results. However, the method still does not provide an insight into RetinaNet. Furthermore, the approach only works in a modular architecture. In the above example, we use RetinaNet as one neural module to detect objects, and we require a second neural module for motor control. We cannot apply this explanation approach to an end-to-end neural architecture that directly maps an input image to a joint configuration because we need suitable samples to train RetinaNet (i.e. bounding boxes of objects), which would not be available for such an end-to-end approach.

As described in Sect. 2.3 Grad-CAM can be applied to any CNN-based neural network to visualize what parts of the input image are relevant to the output. Grad-CAM does not require architectural modifications, so we can directly apply it to an architecture used for end-to-end grasp learning. We adapt the architecture from Kerzel et al. [32], which consists of two convolutional layers (32 filters, 3x3 kernel, ReLU activation), two dense layers (64 units, Sigmoid activation) and an output layer with six neurons (Sigmoid activation), one for each joint of NICO's arm. While the network has only a fraction of the learnable parameters and complexity of state-of-the-art vision networks like RetinaNet, its compactness allows successful training with a limited number of samples.

We create four synthetic datasets; each dataset has 400 samples, consisting of an image of a red-blue grasp object on a table from NICO's perspective and a suitable joint configuration to grasp the object, normalized to [0..1]. Except for the first dataset, one additional distractor object is placed on the table. This distractor object is a small sphere in green, gray and red. Samples are generated by using a genetic inverse kinematics solver [30].

We train the network five times for each dataset with an 80-10-10 train-validation-test split, the Mean Squared Error (MSE) over the joint values as a loss function, and an Adadelta optimizer with an initial learning rate of 0.1, a batch size of 20, and an early-stopping patience of 10.

Table 4 shows the results of the five averaged training runs for each condition. As expected from previous work [27], distractor objects are challenging for such end-to-end approaches; this is reflected in the significant increase in the MSE for all three datasets with distractor objects. We also see that the detrimental effect increases when the distractor object resembles the grasp object. The green distractor caused the smallest increase, and the red distractor object, looking like the grasp object, caused the largest increase. The dataset size of 400 samples is the same as for the initial realization of the approach [32] and works well if no distractor object is present. Still, it can be assumed that more samples are needed for the network to learn to ignore distractors (Fig. 8).

However, this explanation of the network behaviour is gained from extensive experiments with a series of datasets; this method is not applicable to generate explanations during an ongoing interaction. Instead, we are interested in explaining an already trained neural model: We apply Grad-CAM (2.3) to the best-performing model from the dataset without a distractor object. As Grad-CAM only computes a gradient toward a single neuron in the output layer, we choose the first shoulder joint and visualize the relevance of input image regions to see *"which regions of the input image are relevant with regard to computing the values for the first shoulder joint"*, as depicted in Fig. 9. The Grad-CAM visualization shows that both the blue and red parts of the grasping tool contribute substantially to the computation of the joint values. An expert can conclude from this visualization that the network has not learned to ignore task-irrelevant clutter and takes the

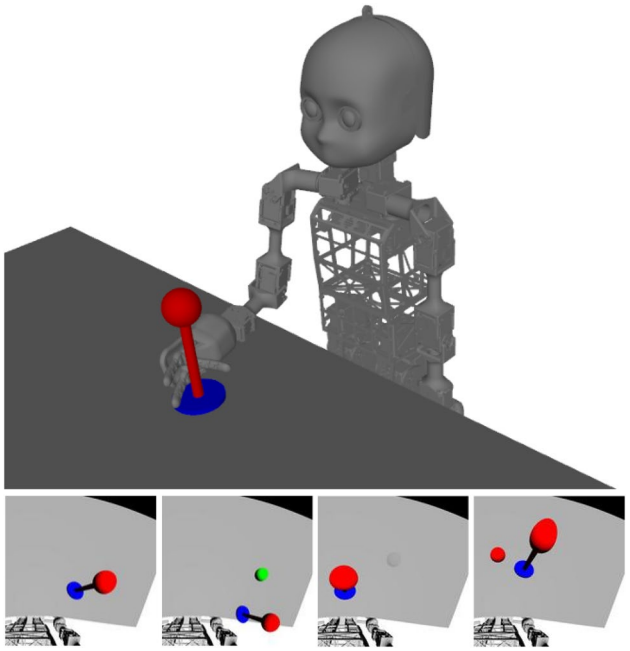

Fig. 8 Top: NICO collects samples in a virtual environment by placing an object and recording its corresponding joint values. Bottom row: Image part of the collected samples without and with distractor objects in different colors (no distractor; green, gray and red distractors)

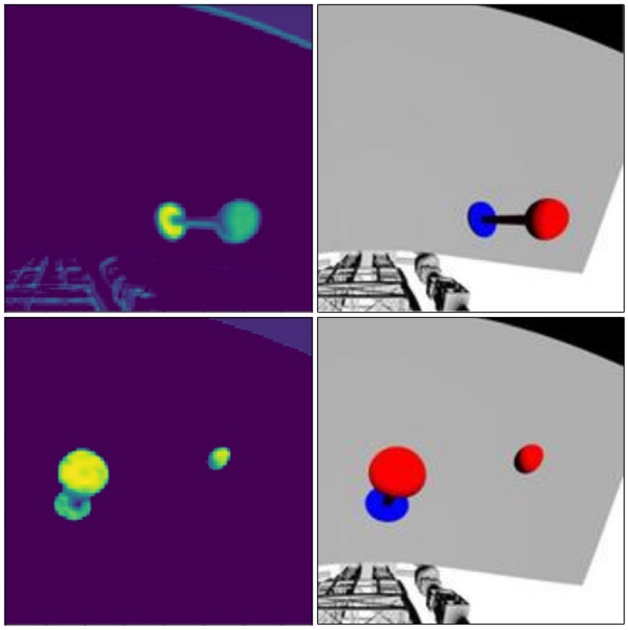

Fig. 9 Grad-CAM visualization of the neural end-to-end grasp approach for the first shoulder joint shows which part of the input image (right side) is relevant to generate the output. The blue and red parts of the grasp object are visible in both Grad-CAM visualizations. The lower Grad-CAM explains why red distractor objects are detrimental for grasping; they show up prominently in the Grad-CAM visualization; the network has not learned to ignore these distractors

Table 4 MSE and SD of training the neural end-to-end grasping approach with different datasets averaged over five training runs. The MSE shows that a distractor object significantly lowers the performance. The effect is more pronounced if the distractor resembles the grasp object

No distractor	Green dist.	Gray dist.	Red dist.
0,03764	0,07312	0,08450	0,09766
$\pm 0,00834$	$\pm 0,00004$	$\pm 0,00029$	$\pm 0,00636$

distractor object into account, thus explaining a possible grasp error. However, for a non-expert, this interpretation might not be obvious. For such a user, a multimodal explanation combining visual and verbal information can be helpful, e.g., the Grad-CAM image could be shown to the user along with the instruction to look for highlighted areas that do not overlap with the grasping object.

Within our framework, this approach serves as an example for generating a multimodal explanation for a neural module that is part of a neuro-robotic architecture; it illustrates that a non-verbal or verbal explanation alone might not always be sufficient to convey the complex neural processing of such a module to a user. Added visualizations, however, can provide an intuitive explanation.

This application of Grad-CAM demonstrates how multimodal explanations, in this case, a visualization along with the verbal or textual description that *highlights relevant parts of the visual input for computing the value of the first shoulder joint* can allow a non-expert user to gain an understanding of how a neural module functions and what might have caused an unwanted behavior.

6 Conclusion

We present a novel an eXplainable Human–Robot Interaction (XHRI) framework for humanoid robot social interaction and collaboration founded on Theory of Mind and hybrid neuro-symbolic architectures. Humans intuitively build a representation of their interaction partner's intentions, knowledge, and reasoning. Based on evidence from literature, it can be assumed that this ability extends to robotic interaction partners. We re-analyze a set of HRI studies conducted on the humanoid robot NICO and show how explanatory communication on different levels (non-verbal cues, verbal and multimodal) improve the perception of a robot in multiple categories.

We argue that the medium of the interaction should be context-dependent, and explanations should use efficient and intuitive communication channels that do not interfere with the actual task at hand. Therefore, our framework suggests using non-verbal social cues such as facial expressions and

gestures to signal the robot's status and issues. If the limited expressiveness of such cues is insufficient, comprehensive verbal explanations can explain the robot's intentions, knowledge, or reasoning. Such explanations are usually sufficient to describe the high-level reasoning and control of a neuro-robotic agent. For explaining neural modules that often realize specialized sensing, actuation or reasoning tasks, it is also desirable to provide a brief verbal explanation. However, sometimes additional visualizations or other ways of communication need to be combined with a verbal message to communicate an explanation for complex neural modules efficiently. Future work in XAI should focus on expressing neural reasoning processes faithfully in natural language and intuitive multimodal representations. In HRI, it is crucial to supplement purely visual explanations with accompanying information using speech or text, to support non-expert users with easy-to-interpret multimodal explanations.

To this end, we provide two example approaches that facilitate explanation for modules in a neuro-robotic architecture. One model enables us to derive verbal explanations for a neural object detector and the second model provides a visualization for a neural grasping approach along with a suggestion on how the visualization can be explained verbally to the user. In future work, we will evaluate the presented neural XHRI framework in further user studies. We hope the proposed XHRI framework will provide a guideline for the formulation of future research and evaluation studies for the integration of XAI and HRI.

Acknowledgements The authors gratefully acknowledge support from the German Research Foundation DFG for the projects CML TRR169, LeCAREbot and IDEAS and the Bundesministerium für Wirtschaft und Klimaschutz BMWK for the project VeriKAS.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

1. Ambsdorf J, Munir A, Wei Y, Degkwitz K, Harms HM, Stannek S, Ahrens K, Becker D, Strahl E, Weber T, et al (2022) Explain yourself! effects of explanations in human–robot interaction. arXiv preprint arXiv:2204.04501
2. Anjomshoae S, Najjar A, Calvaresi D, Främling K (2019) Explainable agents and robots: Results from a systematic literature review. In: 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019, pp 1078–1088. International Foundation for Autonomous Agents and Multiagent Systems
3. Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, García S, Gil-López S, Molina D, Benjamins R et al (2020) Explainable artificial intelligence (xai): concepts, taxonomies, opportunities and challenges toward responsible ai. Inf Fusion 58:82–115
4. Banks J (2020) Theory of mind in social robots: replication of five established human tests. Int J Soc Robot 12(2):403–414
5. Bartneck C, Kulić D, Croft E, Zoghbi S (2009) Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. Int J Soc Robot 1(1):71–81
6. Beik-Mohammadi H, Xirakia N, Abawi F, Barykina I, Chandran K, Nair G, Nguyen C, Speck, D, Alpay T, Griffiths S, Heinrich S, Strahl E, Weber C, Wermter S (2019) Designing a personality-driven robot for a human-robot interaction scenario. In: 2019 IEEE International Conference on Robotics and Automation (ICRA), pp. 4317–4324. Montreal, Canada. https://www2.informatik.uni-hamburg.de/wtm/publications/2019/BXABCNNSAGHSWW19/
7. Calli B, Walsman A, Singh A, Srinivasa S, Abbeel P, Dollar AM (2015) Benchmarking in manipulation research: Using the Yale-CMU-Berkeley object and model set. IEEE Robot Autom Magz 22(3):36–52
8. Churamani N, Anton P, Brügger M, Fließwasser E, Hummel T, Mayer J, Mustafa W, Ng, HG, Nguyen TLC, Nguyen Q, Soll M, Springenberg S, Griffiths S, Heinrich S, Navarro-Guerrero N, Strahl E, Twiefel J, Weber C, Wermter S (2017) The impact of personalisation on human-robot interaction in learning scenarios. In: Proceedings of the Fifth International Conference on Human Agent Interaction, HAI '17, pp 171–180. ACM, ACM, Bielefeld, Germany. https://doi.org/10.1145/3125739.3125756. https://www2.informatik.uni-hamburg.de/wtm/publications/2017/CABFHMMNNNSSGHNSTWW17/
9. Churamani N, Kerzel M, Strahl E, Barros P, Wermter S (2017) Teaching emotion expressions to a human companion robot using deep neural architectures. In: International Joint Conference on Neural Networks (IJCNN), pp 627–634. IEEE, Anchorage, Alaska. https://doi.org/10.1109/IJCNN.2017.7965911. https://www2.informatik.uni-hamburg.de/wtm/publications/2017/CKSBW17/
10. Dautenhahn K, Woods S, Kaouri C, Walters ML, Koay KL, Werry I (2005) What is a robot companion-friend, assistant or butler? In: 2005 IEEE/RSJ international conference on intelligent robots and systems, pp 1192–1197. IEEE
11. Devin S, Alami R (2016) An implemented theory of mind to improve human–robot shared plans execution. In: 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp 319–326. IEEE
12. Dogan FI, Melsión GI, Leite I (2021) Leveraging explainability for comprehending referring expressions in the real world. arXiv:2107.05593
13. Eppe M, Kerzel M, Griffiths S, Ng HG, Wermter S (2017) Combining deep learning for visuomotor coordination with object identification to realize a high-level interface for robot object-picking. In: IEEE-RAS International Conference on Humanoid Robots (Humanoids), pp 612–617. https://www2.informatik.uni-hamburg.de/wtm/publications/2017/EKGNW17/

14. Gelman A, Imbens G (2013) Why ask why? Forward causal inference and reverse causal questions. Tech. rep, National Bureau of Economic Research
15. Goebel R, Chander A, Holzinger K, Lecue F, Akata Z, Stumpf S, Kieseberg P, Holzinger A (2018) Explainable ai: the new 42? In: International cross-domain conference for machine learning and knowledge extraction, pp 295–303. Springer
16. Gray HM, Gray K, Wegner DM (2014) Dimensions of mind perception. Exp Philos 2:88. https://doi.org/10.1093/acprof:osobl/9780199927418.003.0004
17. Greydanus S, Koul A, Dodge J, Fern A (2018) Visualizing and understanding atari agents. In: International conference on machine learning, pp 1792–1801. PMLR
18. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
19. Hegel F, Krach S, Kircher T, Wrede B, Sagerer G (2008) Theory of mind (tom) on robots: a functional neuroimaging study. In: 2008 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp 335–342. IEEE
20. Heider F, Simmel M (1944) An experimental study of apparent behavior. Am J Psychol 57(2):243–259
21. Hellström T, Bensch S (2018) Understandable robots-what, why, and how. Paladyn J Behav Robot 9(1):110–123
22. Iyer R, Li Y, Li H, Lewis M, Sundar R, Sycara K (2018) Transparency and explanation in deep reinforcement learning neural networks. In: Proceedings of the 2018 AAAI/ACM conference on AI, ethics, and society, pp 144–150
23. James S, Davison AJ, Johns E (2017) Transferring end-to-end visuomotor control from simulation to real world for a multi-stage task. In: Conference on robot learning, pp 334–343. PMLR
24. Jung B, Kopp S (2003) Flurmax: an interactive virtual agent for entertaining visitors in a hallway. In: International workshop on intelligent virtual agents, pp 23–26, Springer
25. Juozapaitis Z, Koul A, Fern A, Erwig M, Doshi-Velez F (2019) Explainable reinforcement learning via reward decomposition. In: IJCAI/ECAI Workshop on explainable artificial intelligence
26. Kansky K, Silver T, Mély DA, Eldawy M, Lázaro-Gredilla M, Lou X, Dorfman N, Sidor S, Phoenix S, George D (2017) Schema networks: zero-shot transfer with a generative causal model of intuitive physics. In: International conference on machine learning, pp 1809–1818. PMLR
27. Kerzel M, Abawi F, Eppe M, Wermter S (2020) Enhancing a neurocognitive shared visuomotor model for object identification, localization, and grasping with learning from auxiliary tasks. IEEE transactions on cognitive and developmental systems pp 1–13. https://doi.org/10.1109/TCDS.2020.3028460. https://www2.informatik.uni-hamburg.de/wtm/publications/2020/KAEW20/09211758.pdf
28. Kerzel M, Ng HG, Griffiths S, Wermter S (2017) Effect of a humanoid's active role during learning with embodied dialogue system. In: V.N.T.T.S.W.A.T. Amir Aly Sascha Griffiths (ed) Proceedings of the Workshop: Towards Intelligent Social Robots: Social Cognitive Systems in Smart Environments at the 26th IEEE International Symposium on Robot and Human Interactive Communication. RO-MAN, Lisbon, Portugal. https://www2.informatik.uni-hamburg.de/wtm/publications/2017/KNGW17/
29. Kerzel M, Pekarek-Rosin T, Strahl E, Heinrich S, Wermter S (2020) Teaching nico how to grasp: An empirical study on cross-modal social interaction as a key factor for robots learning from humans. Front Neurorobot 12. https://doi.org/10.3389/fnbot.2020.00028. https://www2.informatik.uni-hamburg.de/wtm/publications/2020/KPSHW20/KPSHW20.pdf
30. Kerzel M, Spisak J, Strahl E, Wermter S (2020) Neuro-genetic visuomotor architecture for robotic grasping. In: M.P.W.S. Farkaš Igor (ed.) Artificial Neural Networks and Machine Learning - ICANN 2020, LNCS, pp. 533–545. Springer. https://doi.org/10.1007/978-3-030-61616-8_43. https://www2.informatik.uni-hamburg.de/wtm/publications/2020/KSSW20/ICANN_2020__Neuro_Genetic_Visuomotor_Framework_Preprint.pdf
31. Kerzel M, Strahl E, Magg S, Navarro-Guerrero, N, Heinrich S, Wermter S (2017) Nico - neuro-inspired companion: A developmental humanoid robot platform for multimodal interaction. In: Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), pp. 113–120. https://www2.informatik.uni-hamburg.de/wtm/publications/2017/KSMNHW17/
32. Kerzel M, Wermter S (2017) Neural end-to-end self-learning of visuomotor skills by environment interaction. In: V.P.V.A. Lintas A. Rovetta S. (ed.) Artificial Neural Networks and Machine Learning - ICANN 2017, Lecture Notes in Computer Science, vol 10613, pp 27–34. Springer, Cham. https://doi.org/10.1007/978-3-319-68600-4_4
33. Kim B, Wattenberg M, Gilmer J, Cai C, Wexler J, Viegas F, et al (2018) Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In: International conference on machine learning, pp. 2668–2677. PMLR
34. Kopp S, Gesellensetter L, Krämer NC, Wachsmuth I (2005) A conversational agent as museum guide–design and evaluation of a real-world application. In: International workshop on intelligent virtual agents, pp 329–343. Springer
35. Krämer NC, Pütten Avd, Eimler S (2012) Human-agent and human-robot interaction theory: similarities to and differences from human–human interaction. In: Human–computer interaction: The agency perspective, pp 215–240. Springer
36. Kruskal WH, Wallis WA (1952) Use of ranks in one-criterion variance analysis. J Am Stat Assoc 47(260):583–621
37. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444
38. Lee Sl, Lau IYm, Kiesler S, Chiu CY (2005) Human mental models of humanoid robots. In: Proceedings of the 2005 IEEE international conference on robotics and automation, pp 2767–2772, IEEE
39. Lescevica M, Ginters E, Mazza R (2013) Unified theory of acceptance and use of technology (UTAUT) for market analysis of FP7 CHOReOS products. Procedia Comput Sci 26(December):51–68. https://doi.org/10.1016/j.procs.2013.12.007
40. Lichtenthäler C, Lorenzy T, Kirsch A (2012) Influence of legibility on perceived safety in a virtual human–robot path crossing task. In: 2012 IEEE RO-MAN: The 21st IEEE international symposium on robot and human interactive communication, pp 676–681, IEEE
41. Lin TY, Goyal P, Girshick R, He K, Dollár P (2017) Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision, pp 2980–2988
42. Madumal P, Miller T, Sonenberg L, Vetere F (2020) Explainable reinforcement learning through a causal lens. In: Proceedings of the AAAI conference on artificial intelligence, vol 34, pp 2493–2500
43. Malle BF (2006) How the mind explains behavior: Folk explanations, meaning, and social interaction. MIT Press, New York
44. Meertens RM, Lion R (2008) Measuring an individual's tendency to take risks: the risk propensity scale. J Appl Soc Psychol 38(6):1506–1520. https://doi.org/10.1111/j.1559-1816.2008.00357.x
45. Miller T (2019) Explanation in artificial intelligence: insights from the social sciences. Artif Intell 267:1–38
46. Miller T, Howe P, Sonenberg L (2017) Explainable ai: beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences. arXiv:1712.00547

47. Mittelstadt B, Russell C, Wachter S (2019) Explaining explanations in ai. In: Proceedings of the conference on fairness, accountability, and transparency, pp 279–288
48. Mnih V, Kavukcuoglu K, Silver D, Graves A, Antonoglou I, Wierstra D, Riedmiller M (2013) Playing atari with deep reinforcement learning. arXiv:1312.5602
49. Nakata T, Sato T, Mori T, Mizoguchi H (1998) Expression of emotion and intention by robot body movement. In: International conference on intelligent autonomous systems 5 (IAS-5), pp 352–359
50. Ng HG, Anton P, Brügger M, Churamani N, Fließwasser E, Hummel T, Mayer J, Mustafa W, Nguyen TLC, Nguyen Q, Soll M, Springenberg S, Griffiths S, Heinrich S, Navarro-Guerrero N, Strahl E, Twiefel J, Weber C, Wermter S (2017) Hey robot, why don't you talk to me? In: Proceedings of the IEEE international symposium on robot and human interactive communication (RO-MAN), pp 728–731. https://doi.org/10.1109/ROMAN.2017.8172383
51. Parisi GI, Kemker R, Part JL, Kanan C, Wermter S (2019) Continual lifelong learning with neural networks: a review. Neural Netw 113:54–71
52. Puri N, Verma S, Gupta P, Kayastha D, Deshmukh S, Krishnamurthy B, Singh S (2019) Explain your move: Understanding agent actions using specific and relevant feature attribution. arXiv:1912.12191
53. Ribeiro MT, Singh S, Guestrin C (2016) "why should I trust you": explaining the predictions of any classifier. arxiv:1602.04938
54. Rist T, Baldes S, Gebhard P, Kipp M, Klesen M, Rist P, Schmitt M (2002) Crosstalk: An interactive installation with animated presentation agents. In: Proceedings of COSIGN, vol. 2. Citeseer
55. Sado F, Loo CK, Kerzel M, Wermter S (2020) Explainable goal-driven agents and robots–a comprehensive review and new framework. arXiv:2004.09705
56. Scassellati B (2002) Theory of mind for a humanoid robot. Auton Robot 12(1):13–24
57. Schrittwieser J, Antonoglou I, Hubert T, Simonyan K, Sifre L, Schmitt S, Guez A, Lockhart E, Hassabis D, Graepel T et al (2020) Mastering atari, go, chess and shogi by planning with a learned model. Nature 588(7839):604–609
58. Selvaraju RR, Das A, Vedantam R, Cogswell M, Parikh D, Batra D (2016) Grad-cam: why did you say that? Visual explanations from deep networks via gradient-based localization. arXiv:1610.02391
59. Sheh RK (2017) Different xai for different hri. In: 2017 AAAI Fall Symposium Series
60. Sheridan TB (2016) Human-robot interaction: status and challenges. Hum Fact 58(4):525–532
61. Simonyan K, Vedaldi A, Zisserman A (2013) Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv:1312.6034
62. Simonyan K, Vedaldi A, Zisserman A (2014) Deep inside convolutional networks: Visualising image classification models and saliency maps. In: In Workshop at International Conference on Learning Representations. Citeseer
63. Takayama L, Dooley D, Ju W (2011) Expressing thought: improving robot readability with animation principles. In: Proceedings of the 6th international conference on Human–robot interaction, pp 69–76
64. Van Hasselt H, Guez A, Silver D (2016) Deep reinforcement learning with double q-learning. In: Proceedings of the AAAI conference on artificial intelligence, vol. 30
65. Voulodimos A, Doulamis N, Doulamis A, Protopapadakis E (2018) Deep learning for computer vision: a brief review. Computational intelligence and neuroscience 2018
66. Williams GC, Deci EL (1996) Internalization of biopsychosocial values by medical students: a test of self-determination theory. J Pers Soc Psychol 70(4):767
67. Wortham RH, Theodorou A (2017) Robot transparency, trust and utility. Connect Sci 29(3):242–248
68. Young T, Hazarika D, Poria S, Cambria E (2018) Recent trends in deep learning based natural language processing. IEEE Comput Intell Magz 13(3):55–75
69. Zahavy T, Ben-Zrihem N, Mannor S (2016) Graying the black box: understanding dqns. In: International conference on machine learning, pp 1899–1908
70. Zhou B, Khosla A, Lapedriza À, Oliva A, Torralba A (2015) Learning deep features for discriminative localization. arXiv:1512.04150
71. Zörner S, Arts E, Vasiljevic B, Srivastava A, Schmalzl F, Mir G, Bhatia K, Strahl E, Peters A, Alpay T, Wermter S (2021) An immersive investment game to study human-robot trust. Front Robot AI 8(June):1–16. https://doi.org/10.3389/frobt.2021.644529