Review



# Towards a Capability Taxonomy for Autonomous Robots in Affective Human–Robot Interaction

Yunjia Sun and Tao Wang *

Posted Date: 25 March 2026

doi: 10.20944/preprints202603.2029.v1

Keywords: human–robot interaction; affective computing; autonomous robots

 doi:10.20944/preprints202603.2029.v1



*Review*

# Towards a Capability Taxonomy for Autonomous Robots in Affective Human–Robot Interaction

**Yunjia Sun and Tao Wang ***

School of Computer Science, Peking University
* Correspondence: wangtao@pku.edu.cn

**Abstract**

Autonomous robots are increasingly integrated into social contexts, making affective human–robot interaction (HRI) critical for their effectiveness and acceptance. However, existing research remains dispersed across domains and techniques, lacking a unified framework to characterize core robotic capabilities. To address this gap, we adopt a capability-oriented perspective and conduct a comprehensive literature review, through which we propose a structured taxonomy of capabilities for robots in affective HRI. The taxonomy comprises five core dimensions: Perception (recognizing human internal states), Strategy (planning responses based on human states and context), Expression (conveying robot lifelikeness and social presence), Sustainability (maintaining effective and reliable operation over time), and Ethics (ensuring behavior within ethical constraints). By organizing diverse research efforts into a structured framework, this taxonomy provides a systematic foundation for designing socially competent robots and guiding future research.

**Keywords:** human–robot interaction; affective computing; autonomous robots

## 1. Introduction

Robots are increasingly deployed in human-centered environments such as education, healthcare, service, companionship, and everyday living, where interaction with humans is frequent, prolonged, and socially situated. In these contexts, interaction often involves not only task coordination but also the interpretation and expression of affective and social signals. To function effectively, robots must be able to understand human internal states, generate contextually appropriate responses, and convey information and intentions through their behavior. Achieving such effectiveness relies on a systematic set of affective HRI-related capabilities. To obtain a systematic understanding of the essential capabilities required for coherent, affect-aware human–robot interaction, a unified perspective is necessary. However, despite significant advancements in specific robot behaviors over the past decade, research contributions remain dispersed across various application domains and technical approaches.

To address the lack of a systematic view, we develop a capability-oriented taxonomy that not only organizes existing research but also clarifies the capabilities enabling robots to interact with humans in affective and socially meaningful ways. This taxonomy provides a framework for understanding the abilities required for effective operation in human-centered environments and for sustaining long-term interaction. It helps researchers situate their work within the broader landscape of affective HRI and provides system designers with a reference for evaluating which capabilities are covered in a robot system and for identifying those that may still be missing.

Based on an extensive literature review, we propose a taxonomy of affective HRI capabilities, organizing them into interrelated categories: Perception, Strategy, Expression, Sustainability, and Ethics. Perception capabilities focus on recognizing and interpreting humans' internal states, such as emotions and intentions. Strategy capabilities concern the generation of adaptive and context-aware interaction policies. Expression capabilities enable robots to convey internal states and social signals

 doi:10.20944/preprints202603.2029.v1



in ways that are both perceivable and meaningful to humans. Sustainability captures capabilities required for long-term interaction. Ethics captures a robot's capability to act in accordance with human ethical norms and to consider potential ethical issues arising from its interactions and use of technology. Our contributions are summerized in three folds. 1) We propose a capability-oriented taxonomy that organizes affective HRI abilities and clarifies their scope with representative examples. 2) We provide a framework that helps researchers situate their work within the broader landscape of affective HRI. 3) We offer designers a reference for evaluating and guiding the development of robots capable of coherent, long-term, affect-aware interaction.

The remainder of this paper is organized as follows. Section 2 reviews related research in affective HRI. Section 3 introduces the proposed taxonomy and discusses the capability categories in detail. Section 4 presents a case study analyzing existing robots using the proposed framework. Finally, Section 5 concludes the paper and discusses future directions.

## 2. Related Work

Many surveys have examined human–robot interaction, covering general topics such as safety and communication, as well as affective aspects such as emotion recognition. In this section, we organize representative surveys into general HRI and affective/emotion-related studies to contextualize our proposed capability taxonomy.

### *2.1. General HRI Surveys*

A number of surveys have examined general aspects of human–robot interaction, such as user experience, safety, communication, *etc.* These studies provide valuable insights into various aspects of HRI.

Several surveys have provided broad thematic overviews of the HRI literature. Lambert et al. [1] survey a decade of research on social robot interaction, covering research topics, application domains, and robot capabilities. Youssef et al. [2] provide an overview of advances in social robotics, including application domains, interaction modalities, robotic platforms, and evaluation methods. Apraiz et al. [3] review 24 studies on UX evaluation in HRI, summarizing commonly evaluated factors, measurement methods, and research gaps. Ostrowski et al. [4] review two decades of HRI research through ethics, equity, and justice perspectives, proposing a framework for inclusive and socially responsible robot design. Robinson et al. [5] review robotic vision in HRI and collaboration, discussing applications, methods, datasets, and evaluation practices.

Some surveys have focused on robots within industrial environments and collaborative contexts. Li et al. [6] review safety methods in industrial human–robot collaboration, categorizing pre- and post-collision strategies and discussing challenges. Rodríguez-Guerra et al. [7] examine challenges in safely deploying collaborative robots and classify HRI developments in industrial settings. Jahanmahin et al. [8] survey human behavior modeling in manufacturing-oriented HRI, covering sensing, prediction, and control strategies.

Research has also investigated how robots communicate with humans. Bonarini [9] reviews communication in HRI, emphasizing the role of linguistic and physical channels in enabling effective interaction. Su et al. [10] summarize multimodal HRI, covering integration of diverse signals including voice, vision, touch, and bio-signals. Nocentini et al. [11] survey behavioral models for social robots, including adaptive behaviors and cognitive architectures.

There is also a body of work focused on HRI taxonomies. Tolmeijer et al. [12] develop a taxonomy of trust-relevant failures and mitigation strategies, categorizing failure types—including design, system, expectation, and user failures—and identifying research gaps for autonomous failure detection and repair. Onnasch and Roesler [13] propose a structured HRI taxonomy that accounts for the human, the robot, the interaction, and the context, providing a framework to enable comparisons across diverse HRI scenarios and identify research gaps. Kim et al. [14] conduct a systematic literature review of robot autonomy in HRI and propose a taxonomy of six distinct forms—operational, intentional, shared,

 doi:10.20944/preprints202603.2029.v1



non-deterministic, cognitive, and physical autonomy—to capture the variety of autonomous behaviors beyond one-dimensional measures.

While these works provide valuable frameworks for general HRI, there remains a lack of a systematic taxonomy specifically focused on robot capabilities within the context of affective HRI. We aim to fill this gap by proposing a capability-centered classification tailored for affective HRI scenarios.

### 2.2. *Affective/Emotion-Related HRI Surveys*

A growing number of surveys have examined affective and emotion-related aspects of human–robot interaction, including emotion recognition, expression, and social-emotional bonding. These studies provide insights into methods, modalities, and models for enabling robots to perceive, respond to, and interact with human emotions.

Several studies have focused on human emotion recognition through visual, vocal, and physiological signals. Spezialetti et al. [15] review advances in emotion recognition, covering emotional models, interaction modalities, and classification strategies. Stock-Homburg [16] survey two decades of research on robotic emotion generation, human recognition, and responses. Cavallo et al. [17] focus on how robots recognize and respond to human emotions using different sensing modalities. Cerqueira [18] examine 72 studies on emotion recognition and robot emotion representation, highlighting user-centered evaluation practices.

Other work emphasizes multimodal and specialized sensing approaches. Ottoni and Filippini et al. [19] review thermal infrared imaging for affective computing, emphasizing its non-intrusive potential. Kovács et al. [20] survey non-verbal communication, gestures, and user personality for enhancing robotic expressivity and trust. Zhao et al. [21] categorize recent robot emotion recognition models by input modality, including vision, language, physiological signals, and multimodal approaches. Gasteiger et al. [22] review prosodic elements in emotional speech, such as tone, loudness, speed, and pauses. Ottoni and Cerqueira [23] also review methods for emotion recognition and robot emotion representation. Savery and Weinberg [24] survey 1,427 publications on robotics and emotion, categorizing approaches into emotional intelligence, emotional models, and implementation.

Previous research has examined high-level emotions. Mitchell and Jeon [25] review 30 studies on attachment in HRI, emphasizing emotional bonds and psychological attachment theories. Hieida and Nagai [26] survey social emotions in robotics, linking psychological and neuroscience findings to robotic recognition, expression, and modeling. de Souza et al. [27] analyze 34 studies on trust and trustworthiness in HRI, highlighting key factors, evaluation approaches, and application domains.

While existing surveys provide valuable overviews of general and affective HRI, they focus on individual dimensions without integrating them into a coherent framework. Our work complements these studies by proposing a unified capability taxonomy and offering a systematic perspective to guide the design of affective robots and future research.

## 3. Taxonomy

To systematically characterize the capabilities required for autonomous robots in affective human–robot interaction, we propose a taxonomy that organizes these capabilities into five interrelated categories: Perception, Strategy, Expression, Sustainability, and Ethics. Figure 1(a) presents our taxonomy, showing subcategories under each major category. The specific rationale and internal structure for each subcategories are further elaborated in the following subsections.

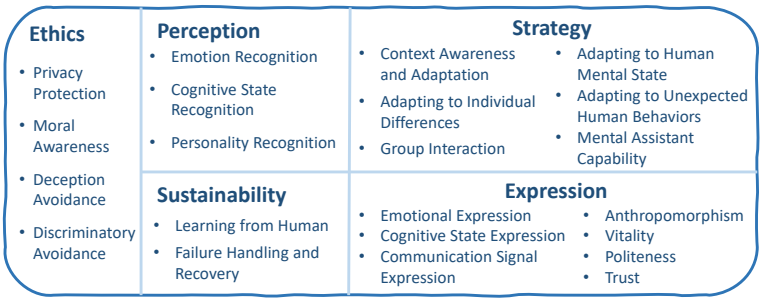
Our taxonomy is conceptually inspired by the sense–plan–act (SPA) paradigm [28,29], in which a robot first senses its environment, then plans appropriate actions, and finally executes them. While originally proposed for system-level task execution, this paradigm provides a useful reference for structuring affective HRI capabilities. Affectie human–robot interaction can be regarded as a specialized robotic task in which the environment includes human partners and their internal states. Consequently, robots require perception, decision-making, and action capabilities specifically tailored to affective interaction.

 doi:10.20944/preprints202603.2029.v1

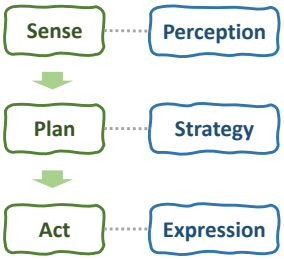


Figure 1(b) illustrates the relation between our taxonomy and the SPA paradigm. In our taxonomy, Perception, Strategy, and Expression correspond to the sense, plan, and act stages, respectively: *Perception* involves sensing and understanding human internal states, *Strategy* corresponds to planning and decision-making for selecting appropriate responses, and *Expression* corresponds to the act of displaying social and affective behaviors in the interaction. Analogous to the SPA paradigm, which follows a sequential process of sense, plan, and act, a robot in affective HRI first perceives human internal states, and then, in conjunction with task requirements and environmental context, formulates an appropriate interaction strategy. Finally, this strategy is realized through the robot's expressive capabilities, which enable the display of corresponding social and affective behaviors during the interaction. In addition, robots should sustain stable and adaptive interaction over time through error recognition, recovery, and continual learning from human partners, which we refer to as *Sustainability*. Last but not least, ethical considerations provide guidance and constraints on the robot's behavior and interaction across all aspects of affective HRI, which we refer to as *Ethics*.

Based on this framework, the subsequent sections introduce each capability category in turn, along with their corresponding sub-categories. It is worth noting that many of these capability classes are themselves extensive research areas, each of which could warrant a dedicated survey. Therefore, rather than aiming for exhaustive coverage, we focus on highlighting a small number of representative works for each category, with the goal of illustrating key ideas and typical approaches observed in the literature.




(a) Taxonomy for Autonomous Robots in Affective Human–Robot Interaction




(b) The sense–plan–act paradigm and its connection with our taxonomy

**Figure 1.** Overview of the proposed taxonomy for affective human–robot interaction capabilities, inspired by the sense-plan-act paradigm. The taxonomy organizes robot capabilities into five core dimensions—Perception, Strategy, Expression, Sustainability and Ethics. Each dimension is further divided into finer-grained capability categories.

### *3.1. Basic Intelligent Capabilities*

Although this paper mainly focuses on affective interaction and the associated capabilities of robots, a set of basic intelligent capabilities remains essential as the foundation supporting effective human–robot interaction. These capabilities underpin a robot's ability to perceive, act, and operate reliably in real-world environments, thereby enabling higher-level affective and social behaviors. In general, such foundational capabilities include, but are not limited to, multimodal perception and information processing, physical manipulation and task execution, autonomous decision-making and control, navigation and environment adaptation, as well as basic communication and coordination mechanisms. In addition, a robot's embodiment and form factor, together with its ability to adapt to different application contexts and team configurations, also play an important role in shaping interaction outcomes.

As these foundational capabilities have been extensively studied in prior HRI and robotics literature, we do not elaborate on them in this paper. Readers interested in comprehensive discussions and systematic taxonomies of basic intelligent capabilities are referred to existing survey and review works, such as [13,30].

 doi:10.20944/preprints202603.2029.v1



### 3.2. *Perception Capability*

Perception capability refers to a robot's ability to infer human internal state, including latent cognitive and affective states. In human–robot interaction (HRI), recognizing human internal states is essential for enabling socially appropriate and adaptive interaction. These internal states include both transient mental states and more stable traits. Mental states further encompass affective states, including emotional states and cognitive states.

Accordingly, we categorize human internal state recognition into three aspects: emotional recognition, cognitive state recognition, and trait recognition. Note that in HRI, recognizing human signals such as gestures and speech is also crucial for effective interaction. However, since these signals are not necessarily related to affective states, they are treated as basic intelligent capabilities and are therefore not discussed in detail here.

#### 3.2.1. Emotional Recognition Capability

Emotional recognition capability refers to the ability to identify the emotions of interaction subjects accurately. Recognizing human emotions is essential in interaction, as emotional states strongly influence human behavior. By perceiving emotional cues, robots can adapt their responses accordingly, enabling more natural and compassionate communication.

To enable emotion recognition in human–robot interaction, researchers have explored various approaches. Mohamed et al. [31] study automatic frustration recognition in human–robot interaction using thermal facial imaging. Focusing on cognitive load– and failure-induced frustration, this work shows that thermal features from key facial regions can effectively capture physiological correlates of frustration, achieving performance comparable to RGB-based facial features. Deng et al. [32] propose a conditional GAN-based framework for facial expression recognition that reduces intra-class variations by jointly learning generative and discriminative representations. By controlling action units and enforcing cycle consistency, the method disentangles expression from identity and other factors, leading to improved FER performance. Mamodiya et al. [33] introduce ECRP, an edge-deployable HRI framework integrating multimodal emotion recognition, context encoding, and reinforcement learning–based task planning for low-latency adaptive behavior on resource-constrained robots.

#### 3.2.2. Cognitive State Estimation Capability

Cognitive state estimation capability refers to the ability to infer the internal cognitive states of interaction subjects, such as intentions, attention, or confusion, from observable behavioral signals. By estimating these states, robots can better understand the needs and ongoing mental processes of human counterparts, enabling more adaptive and proactive interaction.

Intent recognition is one important aspect of cognitive state estimation, referring to inferring the intentions of interaction subjects. By anticipating user needs, robots can proactively assist users and improve interaction efficiency. Intent recognition can be supported by non-verbal cues. Huang and Mutlu [34] propose an anticipatory control framework that infers human task intent from gaze behavior in collaborative settings. By extracting fixation-based gaze features and using an SVM classifier, the robot predicts the human's intended target in advance, enabling proactive action planning and more fluent human–robot collaboration. Ryoo et al. [35] introduce onset cues and cascade histogram-based representations for early activity prediction in robot-centric first-person videos, enabling earlier recognition of human activities for timely robot responses. Intent recognition can also based on text information. Huggins et al. [36] address data scarcity in intent recognition for human–robot interaction using a minimal-data, BERT-based approach. They compare BERT [37] with BiLSTM and logistic regression on public datasets, and further validate the approach on a real-world social robot task involving character strength recognition. Their results highlight the effectiveness of full BERT fine-tuning.

Beyond intention understanding, estimating other cognitive states is also important for effective interaction. For example, detecting user confusion allows robots to adjust dialogue strategies and

 doi:10.20944/preprints202603.2029.v1



provide timely clarification. *Li et al.* [38] investigate interlocutor confusion in situated task-oriented human–robot dialogue. They analyze multimodal cues including emotions, head pose, eye gaze, speech silence duration, and self-reported confusion states, and find significant correlations between these signals and user confusion. Similarly, recognizing user attention is also important for effective human–robot interaction, as it helps robots determine whether users are engaged in the interaction. Chakraborty et al. [39] propose a human–robot interaction system for estimating human attention levels based on visual focus of attention (VFOA). By analyzing facial features, eyeball movement, and gaze using deep learning models, the system enables robots to detect users' attention states and initiate appropriate verbal or visual interaction.

#### 3.2.3. Personality Recognition Capability

Personality recognition capability refers to the ability to accurately identify the personality traits of interaction subjects. It enables robots to develop personalized strategies tailored to the identified personality. This process improves interaction effectiveness, enhances user satisfaction, and fosters more comfortable and engaging experiences.

Researchers have made various attempts to explore personality recognition. Celiktutan et al. [40] present the MHHRI dataset for studying personality and engagement in both HHI and HRI, combining multimodal recordings with self and acquaintance annotations. Salam et al. [41] study engagement prediction in triadic human–human–robot interactions. They show that combining automatically predicted participant personality traits, the robot's personality traits, and nonverbal features improves individual and group engagement classification compared to using nonverbal features alone. The results also indicate that extroverted interactions yield the highest classification performance. Shen et al. [42] propose a framework to infer users' personality traits from habitual visual and vocal behaviors during face-to-face HRI, using questionnaire-based labels to train regression and classification models for improving interaction quality.

### *3.3. Strategy Capability*

In the classical sense–plan–act (SPA) paradigm [28,29], the plan component represents the process through which a robot determines appropriate actions based on perceived information and its internal representation of the environment. Inspired by this perspective, we view strategy capability in affective human–robot interaction as the ability of a robot to determine suitable interaction responses and policies based on its understanding of the situation and the human partner.

As illustrated in Figure 2, a central aspect of this capability is adaptivity. During interaction, robots should be able to adjust their responses according to multiple sources of information, including context information, human mental state, individual difference, and unexpected human behavior. The interaction context and the human internal state can be viewed as extensions of the world model in classical robot architectures. They incorporate socially relevant information beyond the physical environment. Human internal states encompass both mental states and more stable traits; in the context of adaptive interaction, the latter are reflected as individual differences that guide personalized responses. In addition, robots should also be able to react appropriately when human behavior deviates from expected patterns, which we describe as unexpected human behavior in this taxonomy.

 doi:10.20944/preprints202603.2029.v1



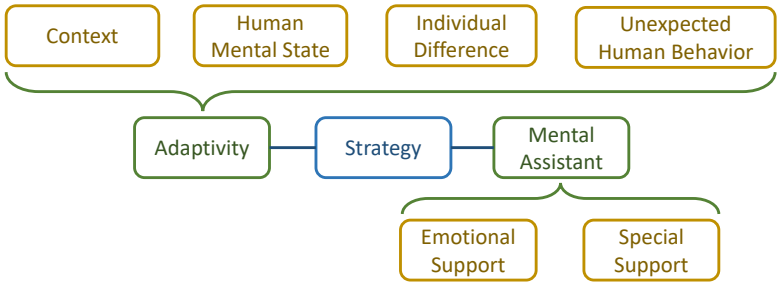



**Figure 2.** Strategic capabilities in affective HRI. These capabilities encompass adaptivity and mental assistant capability. Adaptive capability enables robots to adjust their responses according to context, human mental states, individual differences, and unexpected human behavior. Mental assistant capability encompasses emotional support and special support.

Beyond general adaptive interaction, we identify a specialized dimension of strategic capabilities dedicated to promoting human psychological well-being (Figure 2). An increasing body of research [43–46] explores how robots can actively support mental health or assist interventions for conditions such as autism spectrum disorder and cognitive decline. To capture this line of work, we introduce mental assistant capability, referring to a robot's ability to provide psychological support or facilitate intervention-oriented interactions.

#### 3.3.1. Context Awareness and Adaptation

During interaction, robots should adjust their interaction strategies based on the state of the interaction partner or the task context to enhance their adaptability in real-world scenarios. Hemminghaus and Kopp [47] propose a reinforcement learning–based architecture for adaptive social behavior in assistive robots. Using Q-learning to optimize low-level behaviors from user states, the approach is evaluated in a memory game with the Furhat robot, demonstrating effective adaptation to individual users. Doğan et al. [48] propose an interactive system for resolving object-referential ambiguities in human–robot dialogue through follow-up clarifications, outperforming re-description strategies and improving interaction efficiency and user perceptions. Andriella et al. [49] investigate a Tiago robot with adaptive assistance capabilities in short-term, real-world interactions. Their study demonstrates that the robot can dynamically adjust support based on user performance and task state, improving task outcomes even in one-time, uncontrolled encounters.

#### 3.3.2. Adapting to Human Mental State

In human–robot interaction, a robot's ability to respond appropriately to human mental states is crucial for effective collaboration and socially appropriate interaction. This includes providing empathetic responses and adapting its behavior to better support humans during the interaction. Chen et al. [50] show that affect in child–robot interaction predicts learning outcomes more strongly at fine-grained events than at coarse levels, especially when the robot acts as a tutee, providing cues that can anticipate children's performance. Unhelkar et al. [51] presented CommPlan, a Partially Observable Markov Decision Process framework that explicitly models human intentions as latent states, enabling the robot to decide if, when, and what to communicate—such as sharing its own intention or querying the human's—to improve collaborative efficiency. Devin and Alami [52] proposed a Theory of Mind framework enabling robots to estimate human mental states—including beliefs about goals, plans, and actions—to adapt shared plan execution and provide timely, non-intrusive information during collaboration.

#### 3.3.3. Adapting to Individual Differences

In human–robot interaction, users exhibit substantial individual differences, including personality, preferences, and behavioral tendencies. Effective robots should adapt their interaction strategies

 doi:10.20944/preprints202603.2029.v1



accordingly, tailoring their behaviors to each user to enhance engagement, interaction quality, and overall user experience. Ramachandran et al. [53] investigate personalized break-timing strategies in robot–child tutoring. Using a Nao robot that monitors performance, the system delivers fixed, reward-based, or refocus-based breaks. Results show that personalized strategies lead to greater learning gains and immediate improvements in problem-solving efficiency and accuracy. Li et al. [54] examine how individual differences in trait loneliness affect social robot perception and acceptance. Higher trait loneliness reduces positive anthropomorphic inferences (e.g., warmth, competence) and acceptance, mediated by lower attribution of uniquely human traits such as politeness and organization. Ligthart et al. [55] introduced a memory-based personalization strategy that uses stored interaction history to tailor narrative dialogues, enhancing continuity, closeness, and sustained engagement in long-term child–robot relationships over multiple sessions.

#### 3.3.4. Adapting to Unexpected Human Behavior

Given that human behavior is often boundedly rational, robots need to manage actions that are unexpected or suboptimal. Görür et al. [56] introduce a two-stage Partially Observable Markov Decision Process (POMDP) framework for collaborative robots, predicting human availability, motivation, and capability to infer true help needs, improving collaboration efficiency and interaction naturalness over reactive models. Kwon et al. [57] introduce a risk-aware robot model based on Cumulative Prospect Theory that anticipates suboptimal human behavior, supporting safer and more efficient collaboration than noisy-rational models.

#### 3.3.5. Mental Assistant Capability

Mental assistant capability refers to a robot's ability to provide targeted and adaptive support based on a user's internal states and situational context. This capability focuses on understanding and responding to emotional or psychological needs, enabling the robot to offer appropriate interventions.

Within affective mental assistant capabilities, a key function is supporting users in regulating and coping with negative emotional states, enabling robots to recognize, respond to, and help mitigate negative emotions. Dino et al. [58] present a socially assistive robot that delivers internet-delivered Cognitive Behavioral Therapy (iCBT) through structured conversational interaction, demonstrating its feasibility and acceptance among older adults with depression. Kitt et al. [59] empirically evaluate the role of a socially assistive robot in children's mental health care by examining its stress-buffering effects during a standardized stress task. Although no significant reduction in stress was observed, the study highlights nuanced affective responses and suggests potential benefits for children with higher social anxiety. Laban et al. [60] utilize a social robot in a novel long-term intervention designed to help informal caregivers cope with emotional distress through self-disclosure. Through repeated interactions, caregivers showed increased self-disclosure, improved mood, and reduced loneliness and stress, highlighting the potential of social robots for sustained emotional support.

For individuals with social challenges or other specific conditions—such as autism spectrum disorder (ASD) or dementia—robots can provide targeted support. Ramnauth et al. [43] developed ISTAR, an in-home social robot that delivers realistic workplace-style interruptions to adults with ASD, demonstrating improved interruption tolerance and perceived relevance for enhancing employability through autonomous, role-playing training. Sandygulova et al. [44] investigate how individual differences—such as ASD severity, co-occurring ADHD, verbal ability, and age—affect children's engagement and behavioral outcomes in multi-session robot-assisted autism therapy. Moharana et al. [45] engaged dementia caregivers in co-designing robots that either foster joyful shared moments or alleviate emotional burdens—such as redirecting repetitive questions or acting as the "bad guy"—with roles adapting to dementia stages. Cruz-Sandoval et al. [46] deployed an autonomous social robot to deliver nine weeks of cognitive stimulation therapy to people with dementia, reporting significant reductions in neuropsychiatric symptoms such as agitation, delusions, and euphoria.

 doi:10.20944/preprints202603.2029.v1



### 3.4. Expression Capability

In affective human–robot interaction, expression capability refers to a robot's ability to manifest internal states, convey social signals, and exhibit social legibility and naturalness throughout the interaction. These expressive behaviors not only transmit information but also shape how humans perceive the robot, influencing its perceived lifelikeness and social presence. To capture the roles of these behaviors, we divide the expression capability into three complementary aspects, illustrated in Figure 3, corresponding to shaping the robot's perception as a social agent (*Perceptual Characteristics*), conveying information during interaction (*Information Expression*), and expressing socially meaningful qualities that affect interpersonal perception (*Social Interaction Qualities*).

Perceptual characteristics describe how the robot is perceived as an agent, with anthropomorphism and vitality representing two key perceptual dimensions that contribute to lifelike perception. Information expression focuses on the robot's ability to convey information during interaction, including emotional and cognitive state expression and social signal expression. Finally, social interaction qualities capture the social meanings conveyed through expressive behaviors, which influence how humans interpret the robot's social attitudes during interaction; in this taxonomy, we emphasize politeness and trust as two primary social qualities conveyed through expressive behavior.

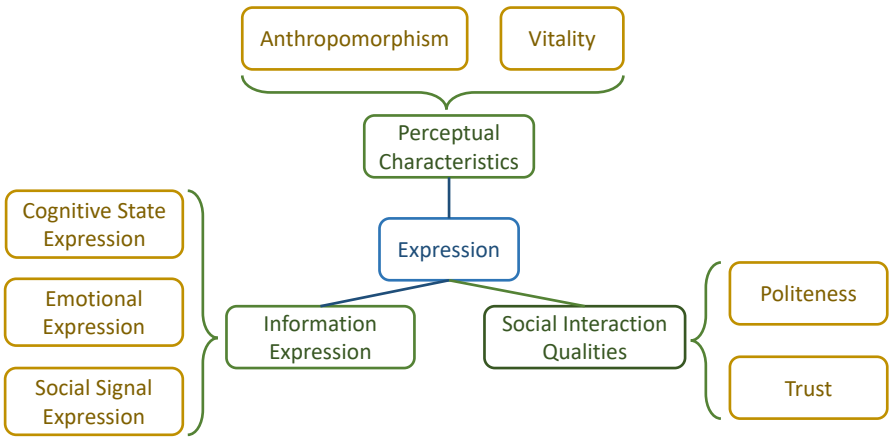



**Figure 3.** Expression-related capabilities in affective HRI. Perceptual characteristics shape how the robot is perceived as a social agent and include anthropomorphism and vitality. Information expression capabilities enable robots to convey internal states and communicative signals. Social interaction qualities capture socially meaningful attributes conveyed through expressive behaviors, including politeness and trust.

#### 3.4.1. Emotional Expression Capability

Emotional expression capability encompasses the ability to effectively translate internal affective states into observable external signals. It allows robots to support natural, socially rich interactions and influence human perception during interaction. Emotions can be expressed through multiple modalities. Velner et al. [61] investigate the effects of intonation in robot speech on conversational naturalness and engagement, finding that humanlike intonation improves perceived naturalness and emotional expressiveness but not turn-taking fluency. Hu and Hoffman [62] demonstrate that dynamically actuated skin textures (e.g., goosebumps, spikes) can effectively convey specific emotions, with texture shape signaling valence and movement frequency indicating arousal. Pelikan et al. [63] analyze how families interpret Cozmo's emotion displays in home settings, showing that "happy" animations advance interaction while "sad" ones act as a "rewind button," prompting users to reassess prior actions. Suguitan et al. [64] proposed MoveAE, a classifying variational autoencoder that enables modification of robot gestures' emotional qualities through latent space arithmetic while preserving recognizability. Block et al. [65] design a human-sized hugging robot that uses softness, warmth, and adaptive haptic responses to deliver emotionally supportive hugs, significantly enhancing perceived naturalness, enjoyability, and friendliness. Au et al. [66] investigate robot-mediated affective touch for

 doi:10.20944/preprints202603.2029.v1



conveying sympathy across U.S. and Japanese users, finding recipients perceive greater social support from robot touch than text, highlighting its role in emotional communication. Dobrosovestnova and Hannibal [67] argue for including ambivalent emotions like teachers' disappointment in robotic tutors, framing it as a care-based, affiliative pedagogical strategy rather than purely negative affect.

To systematically study robotic emotional expression, standardized evaluation tools are required to quantify how clearly and distinctly emotions are perceived by human users. Coyne et al. [68] propose the Geneva Emotion Wheel as a systematic tool to evaluate the perceived emotional expressiveness of robots, enabling quantitative analysis of emotion distinctness and clarity in HRI.

#### 3.4.2. Cognitive State Expression Capability

Cognitive state expression refers to how robots convey internal mental states—such as attention, intention, or curiosity—through observable behaviors. By signaling these internal cognitive processes, robots enable humans to better perceive and interpret the robot's state, thereby enhancing interaction transparency and effectiveness. Gordon et al. [69] demonstrate that children exhibit increased exploratory behavior and uncertainty seeking after interacting with a curious social robot, suggesting curiosity can be socially elicited in human–robot interaction. Zhang et al. [70] examined how curiosity-driven off-task robot behaviors influence human perception, showing that while such actions were interpreted as curious, they reduced perceived competence—unless accompanied by explanatory cues. Aliasghari et al. [71] demonstrate that a humanoid's gaze allocation and arm motion smoothness significantly shape human teachers' perceptions of its confidence, eagerness to learn, and task attention—key cognitive states in learning interactions. Briggs et al. [72] integrate an attention-driven cognitive framework (ARCADIA) with a robotic architecture (DIARC) to generate dynamic, human-like gaze that reflects the robot's internal attentional focus during object learning tasks.

#### 3.4.3. Social Signal Expression Capability

The capability for Social Signal Expression enables a robot to convey social cues through observable behaviors. It helps humans interpret the robot's actions, understand its social role, and engage appropriately during interaction.

Gaze is considered an important form of social expression. Terzioğlu et al. [73] enhance collaborative robots with gaze and breathing cues inspired by animation principles, showing that such social signals improve perceived likeability, animacy, and intention communication. Pereira et al. [74] evaluate a responsive joint attention gaze system in HRI across different contexts, showing that gaze behaviors enhance perceived social presence in adults and external observers but not in children.

Gestures enable robots to express meaning and affect through movement. De Wit et al. [75] examine how varied iconic gestures by a robot affect children's engagement and second-language learning, showing that gesture use—whether repeated or varied—increases social engagement but not vocabulary gains. Roy et al. [76] propose GPT-Driven Gestures, an LLM-based framework that autonomously generates expressive robot gestures, showing human-comparable clarity and improved state recognition in human–robot interaction.

Multimodal communication signals, including verbal cues and expressive behaviors, allow robots to convey social roles. Song et al. [77] design and validate evaluative versus non-evaluative robot roles—using verbal cues, gaze, and appearance—to influence children's motivation during musical practice, demonstrating how social role framing shapes interaction dynamics. Song et al. [78] investigate how three levels of operator presence—Wizard of Oz, costume, and telepresence—affect customer behavior in a supermarket, showing that moderate social presence (costume form) yields optimal engagement and task performance.

#### 3.4.4. Politeness

The ability to exhibit politeness determines whether a robot can regulate social norms and interactional appropriateness. It helps ensure interactions are socially appropriate and comfortable for

 doi:10.20944/preprints202603.2029.v1



users. Kato et al. [79] design a polite approaching behavior for robots by modeling human service staff, enabling the robot to initiate interactions only when visitors show intent to engage, thereby reducing intrusiveness in public spaces. Jackson et al. [80] investigate how robot and human gender influence perceptions of robotic command refusals, finding that politeness norms in noncompliance are significantly shaped by gender alignment and stereotypes. Rea et al. [81] demonstrate that impolite robot encouragement during exercise increases participants' effort and competitiveness, despite lower likability, challenging the assumption that robots should always be polite.

#### 3.4.5. Trust

The ability to establish trust reflects user confidence in robot reliability and competence. Natarajan and Gombolay [82] found that robot behavior and perceived anthropomorphism—not physical embodiment—most strongly influence trust and compliance, with apologetic feedback increasing trust and accountability reducing over-reliance after errors. Herse et al. [83] demonstrated that an initial behavioral measure of trust—assessed via an adapted Trust Game—significantly predicts user decision-making and task performance in human-agent collaboration, with correct assistance enabling recovery on difficult tasks.

Some studies on trust repair will also be discussed in Section 3.5.2 *Failure Handling and Recovery* [12,84], as robot-induced failures can undermine users' trust, necessitating appropriate responses or actions to restore trust.

#### 3.4.6. Anthropomorphism

Anthropomorphism is an aspect of robot expressive capability, reflecting the extent to which users perceive the robot as human-like. Focusing on anthropomorphic design, Bryant et al. [85] investigate how robot gendering and occupational gender-roles influence human trust, concluding that perceived occupational competency—rather than gendered traits—is the primary driver of trust. Torre et al. [86] show that people are more likely to yield to robots perceived as teleoperated and keep closer distances to highly anthropomorphic robots, highlighting how perceived autonomy and human-likeness shape navigation behavior in social dilemmas. Hover et al. [87] analyze online commentary on robots of varying humanlikeness and gender, finding that highly humanlike robots elicit more negative, uncanny, and sexualized responses—especially when female—highlighting risks of anthropomorphic design.

#### 3.4.7. Vitality

The ability to exhibit vitality determines the extent to which a robot appears alive or lively. Iizawa and Yamanaka [88] present "Face on a Globe," a non-anthropomorphic spherical robot that uses smooth geometric deformation and autonomous orientation to evoke lifelikeness while preserving artificiality. Löffler et al. [89] demonstrate an uncanny valley effect in zoomorphic robots, showing that likeability follows a U-shaped curve with animal likeness—highlighting the perceptual tension between artificiality and biological realism.

### *3.5. Sustainability Capability*

The capability for sustainability enables a robot to maintain effective and reliable operation over time, including learning from human input, recovering from failures, and progressively improving performance, which is essential for long-term interaction and user trust. Figure 4 shows the organization of sustainability-related capabilities. Conceptually, sustainability in affective HRI can be divided into two complementary aspects: leveraging human guidance to improve learning and handling failures to maintain reliable operation. Learning from human captures the robot's ability to acquire knowledge or refine behavior based on human input. This includes human correction, human preference, and broader human feedback, which together enable continuous adaptation and personalization. Failure handling in this work refers specifically to capabilities relevant to affective HRI, focusing on detecting and responding to failures in a way that supports socially aware interaction and maintains user trust. Failure handling can be further divided into failure detection and failure response. Failure detection

 doi:10.20944/preprints202603.2029.v1



involves identifying errors or unexpected events by observing human reactions to the robot's behavior. Failure response refers to strategies for responding to detected failures during interaction, including providing explanations and repairing human perceptions of the robot in order to restore trust and maintain effective interaction.

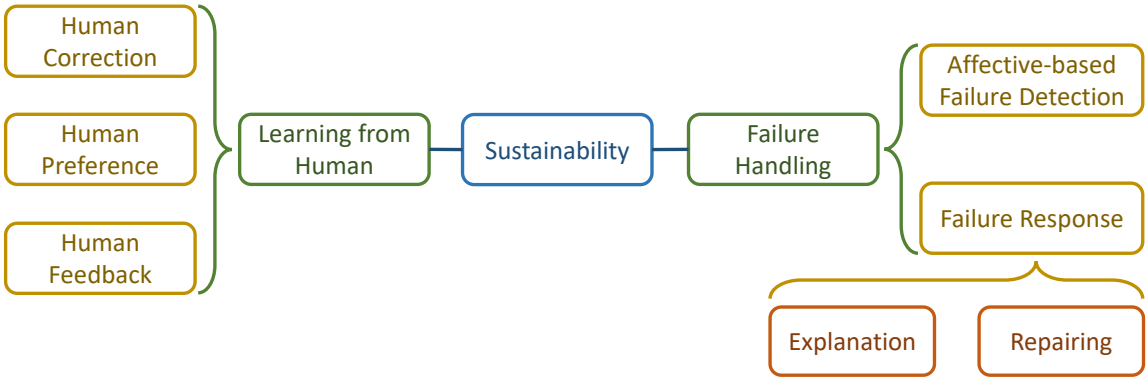



**Figure 4.** Hierarchical organization of sustainability-related capabilities in affective HRI. It is divided into two primary aspects: learning from human input, including correction, preference, and feedback, and failure handling, which encompasses detecting errors and responding to them (including providing explanations and repairing human perceptions).

#### 3.5.1. Learning from Human

A critical capability for robots to achieve sustainability in affective HRI is the ability to learn directly from humans. This enables autonomous humanoids to adapt their responses and progressively refine their affective skills through imitation, feedback, or demonstration, fostering more natural engagement.

Robots can possess the ability to learn from human corrections. Bajcsy et al. [90] propose a one-at-a-time learning approach where robots infer which single task feature a human intends to correct during physical interaction, reducing unintended learning compared to updating all features simultaneously. Losey et al. [91] propose a framework for robots to learn task objectives online from physical human-robot interaction, treating human physical corrections not as disturbances to be rejected, but as informative signals for real-time model updates.

The capability to learn human preferences is an aspect of the robot's learning-from-human capability. Basu et al. [92] enhance comparison-based reward learning by augmenting preference queries with feature queries that ask users which feature explains their preference, enabling faster and more accurate acquisition of human intent. Chen et al. [93] propose Multi-Style Reward Distillation to jointly infer a shared task reward and individual strategy rewards from heterogeneous human demonstrations, effectively disentangling goals from strategic preferences in imitation learning. Bobu et al. [94] propose LESS, a probabilistic human behavior model that accounts for trajectory similarity to better infer human preferences, improving prediction accuracy and inference robustness over the standard Boltzmann rationality model in continuous action spaces.

The ability to learn from human feedback is also crucial. McQuillin et al. [95] present an adaptive robo-waiter that learns socially appropriate behaviors through real-time implicit (facial affect) or explicit (verbal) user feedback, showing improved perceived sociability and positioning appropriateness over static policies. Moreira et al. [96] proposed an interactive deep reinforcement learning (IDeepRL) approach that uses interactive feedback from human or artificial trainers to accelerate the learning of domestic robotic tasks. Their results demonstrate that such feedback significantly reduces learning time and errors compared to standard autonomous deep reinforcement learning.

Beyond single-source learning, some researchers have explored approaches that incorporate both correction and feedback. Celemin and Kober [97] introduced an interactive imitation learning approach that enables robots to proactively request corrective or evaluative feedback based on estimated epistemic and aleatoric uncertainty. This framework allows the robot to handle demonstration ambiguity

 doi:10.20944/preprints202603.2029.v1



and improve data efficiency by strategically leveraging human input when its own knowledge is insufficient. Chisari et al. [98] proposed the CEILing framework, which enables robots to learn complex manipulation tasks by asynchronously integrating both corrective and evaluative human feedback into a stochastic policy. Their approach significantly improves learning efficiency, allowing robots to train directly from raw visual observations in less than one hour of real-world interaction.

To be effective in everyday settings, robots should be able to learn from non-expert users, not just experts. Van Waveren et al. [99] propose a shield-based policy repair method that uses corrective feedback from non-experts after robot failures to refine actions, substitute objects, or forbid unsafe behaviors, improving retraining efficiency in reinforcement learning. Schrum et al. [100] propose MIND MELD, a personalized meta-learning framework that infers user-specific feedback styles via variational embeddings to correct suboptimal human labels in robot-centric imitation learning, improving task performance and user trust.

#### 3.5.2. Failure Handling and Recovery

In the context of affective HRI, failure handling encompasses both the identification of breakdowns via human affective cues and the management of socio-affective errors. We focus on capabilities for detecting and recovering from failures that directly impact the affective loop or leverage human emotional responses as diagnostic signals. Consequently, purely technical fault tolerance and diagnostic mechanisms that do not directly involve human emotional dimensions are excluded from this discussion.

Failure detection is a critical capability that enables robots to recognize breakdowns or deviations in task execution, interaction processes, or system states. By detecting such failures in a timely manner, robots can maintain coherent task and environmental awareness and better cope with unforeseen situations during interaction. In affective human–robot interaction, human reactions can serve as cues for failure recognition. Trung et al. [101] propose an automatic error detection approach in human–robot interaction by classifying human head and shoulder movements. Using RGB-D data collected during interactions involving social norm violations and technical failures, they show that human motion cues can effectively indicate error situations, with performance varying across classifiers and user familiarity. Kontogiorgos et al. [102] investigate users' behavioral responses to robot conversational failures by manipulating embodiment (a human-like robot versus a smart speaker) and failure severity. They train a random forest classifier and compare its performance with human annotations, showing that embodiment significantly shapes gaze and speech cues used for failure detection. Loureiro et al. [103] developed a pipeline for robots to detect and classify social norm violations and technical failures by monitoring human non-verbal social signals, such as facial expressions and eye gaze, through the robot's onboard camera. Their approach combines these social cues with robot context logs to achieve accurate error recognition.

In human–robot interaction, the occurrence of failures not only affects task execution but also shapes how humans perceive the robot. Therefore, a robot's ability to respond appropriately to failures is crucial for maintaining interaction quality and positive user perceptions. Some work focus on the explanation of failures. Kwon et al. [104] propose an optimization-based method for robots to generate expressive motions conveying incapability. These motions communicate task goals and failure causes, improving user understanding, robot perception, and willingness to collaborate. Stange and Kopp [105] investigate how self-explanations by social robots affect the perception of surprising or undesirable intentional behaviors. Explanations based on folk-psychology strategies improved understandability and desirability, with causal explanations proving most effective. Das et al. [106] develop natural language explanations for robot fault recovery for non-expert users, integrating recent action history and environmental context. Their encoder–decoder model autonomously generates explanations that generalize to new environments and match hand-scripted methods in supporting failure and solution identification. Other works focus on repairing human perceptions of the robot. Tolmeijer et al. [12] propose a taxonomy of four trust-relevant HRI failure types—Design, System, Expectation, and User—and corresponding mitigation strategies such as explanation, apology, and

 doi:10.20944/preprints202603.2029.v1



training. Green et al. [107] found affiliative humor improved bystanders' competence perceptions, while less robot-experienced users favored successful robots, highlighting perception-oriented failure responses in HRI. Van der Hoorn et al. [84] found that robots blaming themselves for collaborative failures—especially when incorrect—were perceived as more trustworthy, friendlier, and more desirable for future collaboration than those blaming human partners.

### *3.6. Ethical Capability*

In affective human–robot interaction, ethical capability is particularly important, as various capabilities in perception, strategy, expression, and sustainability can give rise to ethical concerns. Robots therefore require the ability to act within ethical constraints, which we categorize into privacy protection, moral awareness, deception avoidance, and discriminatory avoidance.

#### 3.6.1. Privacy Protection Capability

Privacy protection capability denotes a robot's ability to appropriately handle information arising during interaction. This capability supports the maintenance of user trust by enabling context-aware management of informational boundaries in affective interactions. Tang et al. [108] proposed CONFIDANT, a privacy controller for social robots that uses conversational metadata (e.g., sentiment, topic, relationships) to infer privacy boundaries and regulate information sharing, significantly improving perceived trustworthiness and social awareness in human-robot interaction. Dorafshanian et al. [109] present a differential privacy library for social robots that enables nontechnical users to securely share statistical data. By incorporating risk thresholds and disclosure impact assessment, the approach mitigates privacy risks associated with large-scale personal data collection in social robotics. Aryania et al. [110] examine how design transparency affects trust and data-sharing behavior in public human–robot interactions. Through a user study with a social robot, they show that higher transparency and data sensitivity awareness lead to more cautious data-sharing decisions, despite limited impact on user trust.

#### 3.6.2. Moral Awareness Capability

Moral awareness capability enables a robot to recognize the ethical and normative dimensions of an interaction. This capability involves identifying morally relevant situations and aligning affective responses with established social expectations. Jackson and Williams [111] demonstrated that language-capable robots generating clarification requests for morally impermissible commands can inadvertently signal willingness to violate norms, thereby miscommunicating intentions and weakening human adherence to those moral norms. Wen et al. [112] introduced a computational framework for robots to generate norm-violation responses grounded in Confucian role ethics, showing that explanations incorporating relational roles and context enhance trust, understanding, and perceived intelligence depending on the social role enacted.

#### 3.6.3. Deception Avoidance Capability

Deception avoidance capability refers to a robot's ability to maintain appropriate alignment between its expressed behaviors and the expectations it evokes. In affective human–robot interaction, this capability helps prevent misunderstanding and misplaced trust arising from emotional expressiveness or social cues. Lacey and Caudwell [113] argue that the "cute" aesthetic in home robots functions as a dark pattern by exploiting affective responses to obscure data collection practices, thereby undermining user agency and long-term privacy considerations. Winkle et al. [114] applied ethical risk assessment to socially assistive robots, finding that anthropomorphic behaviors—though potentially deceptive—are generally acceptable and effective, advocating for customizable anthropomorphism rather than its outright avoidance. Sharkey and Sharkey [115] highlight that deception in social robotics can occur with or without intent, leading to overestimated robot capabilities and misplaced trust. They argue that harmful impacts on individuals and society warrant ethical scrutiny, responsibility allocation, and regulatory frameworks.

 doi:10.20944/preprints202603.2029.v1



#### 3.6.4. Discriminatory Avoidance Capability

Discriminatory avoidance capability denotes a robot's ability to support equitable interaction across individuals and social groups. It concerns the regulation of affective perception and behavior to reduce bias and ensure fair treatment in diverse human–robot interaction contexts. Winkle et al. [116] demonstrated across U.S., Swedish, and Japanese samples that female-presenting robots responding to sexist abuse with rationale-based, norm-breaking replies significantly enhanced credibility without reinforcing harmful gender stereotypes. Trainer et al. [117] examine affinity bias in human–robot interaction, showing that participants' trust and teammate selection are influenced by avatar gender and skin tone. Even when competency dominates choices, perceived similarity leads to biased preferences, highlighting potential risks of discriminatory behavior in HRI. Barfield [118] investigates inclusive design for social robots, emphasizing the accommodation of diverse user groups across gender, culture, and religion. Drawing on Social Identity Theory, the work proposes guidelines for human–robot interfaces that promote equity, diversity, and inclusion in interactive contexts. Londoño et al. [119] survey fairness in robot learning, identifying sources of bias and resulting discrimination. The work spans technical, ethical, and legal perspectives, highlighting unfair outcomes and strategies to mitigate bias, providing a foundation for equitable and responsible robot learning. Fossa and Sucameli [120] examines the ethical implications of intentionally embedding gender cues in conversational agents, analyzing how such design choices can trigger social biases. The work evaluates strategies to minimize discriminatory effects, informing socially responsible design in human–robot interaction.

## 4. Case Study on Existing Robots

In this section, we apply the proposed taxonomy to three representative robots: Optimus, Pepper, and Erica. These robots were selected for their diversity in design and function. Optimus is a humanoid robot primarily oriented toward physical task execution, with capabilities such as locomotion, object manipulation, and interaction with physical environments. Pepper is a social robot designed for human interaction, with capabilities focused on communication, emotion perception, and expressive behaviors. Erica is a highly humanlike android specialized in conversational interaction and expressive communication. Using the taxonomy, we highlight their capabilities across Perception, Strategy, Expression, Sustainability, and Ethics. Relevant information about these robots was obtained from the *RobotsGuide* website[1].

### *4.1. Optimus*

Optimus, also known as Tesla Bot, is a humanoid robot developed by Tesla and designed as a general-purpose robotic assistant. It aims to perform repetitive, dangerous, or undesirable tasks for humans, particularly in industrial and everyday environments. Optimus can demonstrate capabilities such as autonomous navigation, obstacle avoidance, and object manipulation to assist with physical tasks.

Based on videos of Optimus shown on the *RobotsGuide* website[2], its affective human–robot interaction capabilities can be categorized according to our taxonomy. In terms of *Perception*, Optimus can recognize human intentions, including aggressive intent and the desire for specific objects, demonstrating its *Cognitive State Recognition capability*. In *Strategy*, it can respond to a person's desire for an object by handing it over, reflecting its ability to *Adapt to Human Mental States*. Regarding *Expression*, Optimus reciprocates greetings from humans, showing *Politeness* in interaction.

### *4.2. Pepper*

Pepper is a humanoid robot developed by SoftBank Robotics, designed for social interaction. It can perceive its environment and interact with humans through speech, vision, and touch, and is

[1] https://robotsguide.com/
[2] https://youtu.be/DrNcXgoFv20

 doi:10.20944/preprints202603.2029.v1



capable of recognizing basic emotions and engaging users through expressive behaviors. Pepper is widely used in retail, public spaces, and research settings to facilitate human–robot interaction.

Based on Pepper's promotional videos, *Pepper size scanner*[3] and *Pepper SPECIAL MOVIE "Let the future begin."*[4], its affective human–robot interaction capabilities can be categorized according to our taxonomy. In terms of *Perception*, Pepper is able to detect when a person is unhappy, demonstrating *Emotion Recognition* capability. In *Strategy*, it can comfort individuals who are feeling unhappy, reflecting its *Mental Assistant* capability. Regarding *Expression*, Pepper conveys information and emotions through changes in eye color, voice intonation, and body posture, reflecting *Emotional Expression* capability. It can also express affirmation through gestures such as nodding, reflecting *Social Signal Expression* capability. Moreover, Pepper maintains a continuous and lively presence during interaction rather than remaining inactive when not performing explicit tasks, which reflects its *Vitality*. When a person behaves aggressively toward Pepper, it can respond with appropriate feedback to comfort the individual, demonstrating *Politeness* and helping to establish a sense of *Trust*.

### *4.3. Erica*

Erica is a highly realistic humanoid android developed by researchers at Osaka University, Kyoto University, and the Advanced Telecommunications Research Institute (ATR) as a research platform for studying human–robot interaction. Designed to support natural conversation with humans, Erica can understand spoken language, generate speech with a synthesized human-like voice, and display a variety of facial expressions and head movements to support social interaction. She has been widely used in research exploring conversational interaction and humanlike presence in social robots.

Based on videos of Erica shown on the *RobotsGuide* website, her affective human–robot interaction capabilities can also be analyzed according to our taxonomy. In terms of *Perception*, Erica is able to recognize human laughter (demonstrated in the video *Shared Laughter with ERICA*[5]), reflecting her *Emotion Recognition* capability. After detecting the user's laughter, she responds by laughing along, demonstrating social responsiveness. This behavior can be categorized under *Strategy*, specifically *Context Awareness and Adaptation*, as she produces an appropriate response based on the user's state. A notable strength of Erica lies in her expressive ability. With a highly humanlike face, she is capable of conveying emotional information through rich facial expressions and gaze behaviors, reflecting *Emotional Expression* and *Cognitive State Expression* capabilities. Furthermore, Erica's extremely humanlike appearance also reflects a high level of *Anthropomorphism*.

## 5. Conclusion

We propose a unified capability taxonomy for autonomous robots in affective human–robot interaction (HRI) to address the dispersion of existing research. The taxonomy is structured around five dimensions—Perception, Strategy, Expression, Sustainability, and Ethics. The taxonomy, which organizes diverse research efforts into a coherent framework, provides a systematic reference for robot design and future studies. Overall, it aims to support the development of socially competent robots capable of natural, trustworthy, and meaningful interactions with humans.

While the current work presents a capability taxonomy that organizes key functions for affective HRI, it remains a conceptual framework. Future work could extend this taxonomy into a comprehensive evaluation standard, not only defining the relevant categories but also specifying criteria for assessing performance within each category. Such an extension would provide practical guidance for designing, benchmarking, and comparing affective robotic systems, moving from a descriptive framework toward actionable assessment and design support.

[3] https://youtu.be/u9TiVa-_af4
[4] https://youtu.be/pKEeZvXSyQU
[5] https://youtu.be/6tMiWog4l00

 doi:10.20944/preprints202603.2029.v1



**Author Contributions:** Conceptualization, Yunjia Sun and Tao Wang; methodology, Yunjia Sun; investigation, Yunjia Sun and Tao Wang; writing—original draft preparation, Yunjia Sun; writing—review and editing, Yunjia Sun and Tao Wang; All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the Joint Laboratory of Peking University and Beijing Innovation Center of Humanoid Robotics Co., Ltd. on Affective Intelligence Application project.

**Data Availability Statement:** Data sharing is not applicable.

**Acknowledgments:** During the preparation of this manuscript/study, the author(s) used ChatGPT, Gemini, and doubao for the purposes of polishing the language. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

**Conflicts of Interest:** The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

## References

1. Lambert, A.; Norouzi, N.; Bruder, G.; Welch, G. A Systematic Review of Ten Years of Research on Human Interaction with Social Robots. *International Journal of Human–Computer Interaction* **2020**, *36*, 1804–1817. https://doi.org/10.1080/10447318.2020.1801172.
2. Youssef, K.; Said, S.; Alkork, S.; Beyrouthy, T. A Survey on Recent Advances in Social Robotics. *Robotics* **2022**, *11*. https://doi.org/10.3390/robotics11040075.
3. Apraiz, A.; Lasa, G.; Mazmela, M. Evaluation of User Experience in Human–Robot Interaction: A Systematic Literature Review. *International Journal of Social Robotics* **2023**, *15*, 187–210. https://doi.org/10.1007/s12369-022-00957-z.
4. Ostrowski, A.K.; Walker, R.; Das, M.; Yang, M.; Breazea, C.; Park, H.W.; Verma, A. Ethics, Equity, & Justice in Human-Robot Interaction: A Review and Future Directions. In Proceedings of the 2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), 2022, pp. 969–976. https://doi.org/10.1109/RO-MAN53752.2022.9900805.
5. Robinson, N.; Tidd, B.; Campbell, D.; Kulić, D.; Corke, P. Robotic Vision for Human-Robot Interaction and Collaboration: A Survey and Systematic Review. *J. Hum.-Robot Interact.* **2023**, *12*. https://doi.org/10.1145/3570731.
6. Li, W.; Hu, Y.; Zhou, Y.; Pham, D.T. Safe human–robot collaboration for industrial settings: a survey. *Journal of Intelligent Manufacturing* **2024**, *35*, 2235–2261. https://doi.org/10.1007/s10845-023-02159-4.
7. Rodríguez-Guerra, D.; Sorrosal, G.; Cabanes, I.; Calleja, C. Human-Robot Interaction Review: Challenges and Solutions for Modern Industrial Environments. *IEEE Access* **2021**, *9*, 108557–108578. https://doi.org/10.1109/ACCESS.2021.3099287.
8. Jahanmahin, R.; Masoud, S.; Rickli, J.; Djuric, A. Human-robot interactions in manufacturing: A survey of human behavior modeling. *Robotics and Computer-Integrated Manufacturing* **2022**, *78*, 102404. https://doi.org/https://doi.org/10.1016/j.rcim.2022.102404.
9. Bonarini, A. Communication in Human–Robot Interaction. *Current Robotics Reports* **2020**, *1*, 279–285. https://doi.org/10.1007/s43154-020-00026-1.
10. Su, H.; Qi, W.; Chen, J.; Yang, C.; Sandoval, J.; Laribi, M.A. Recent advancements in multimodal human–robot interaction. *Frontiers in Neurorobotics* **2023**, *17*, 1084000. https://doi.org/10.3389/fnbot.2023.1084000.
11. Nocentini, O.; Fiorini, L.; Acerbi, G.; Sorrentino, A.; Mancioppi, G.; Cavallo, F. A Survey of Behavioral Models for Social Robots. *Robotics* **2019**, *8*. https://doi.org/10.3390/robotics8030054.
12. Tolmeijer, S.; Weiss, A.; Hanheide, M.; Lindner, F.; Powers, T.M.; Dixon, C.; Tielman, M.L. Taxonomy of Trust-Relevant Failures and Mitigation Strategies. In Proceedings of the 2020 15th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2020, pp. 3–12.
13. Onnasch, L.; Roesler, E. A Taxonomy to Structure and Analyze Human–Robot Interaction. *International Journal of Social Robotics* **2021**, *13*, 833–849. https://doi.org/10.1007/s12369-020-00666-5.
14. Kim, S.; Anthis, J.R.; Sebo, S. A Taxonomy of Robot Autonomy for Human-Robot Interaction. In Proceedings of the 2024 19th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2024, pp. 381–393.
15. Spezialetti, M.; Placidi, G.; Rossi, S. Emotion Recognition for Human–Robot Interaction: Recent Advances and Future Perspectives. *Frontiers in Robotics and AI* **2020**, *7*, 532279. https://doi.org/10.3389/frobt.2020.532279.

 doi:10.20944/preprints202603.2029.v1




16. Stock-Homburg, R. Survey of Emotions in Human–Robot Interactions: Perspectives from Robotic Psychology on 20 Years of Research. *International Journal of Social Robotics* **2022**, *14*, 389–411. https://doi.org/10.1007/s12369-021-00778-6.
17. Cavallo, F.; Semeraro, F.; Fiorini, L.; Magyar, G.; Sinčák, P.; Dario, P. Emotion Modelling for Social Robotics Applications: A Review. *Journal of Bionic Engineering* **2018**, *15*, 185–203. https://doi.org/10.1007/s42235-018-0015-y.
18. Ottoni, L.T.C.; Cerqueira, J.d.J.F. A Systematic Review of Human–Robot Interaction: The Use of Emotions and the Evaluation of Their Performance. *International Journal of Social Robotics* **2024**, *16*, 2169–2188. https://doi.org/10.1007/s12369-024-01178-2.
19. Filippini, C.; Perpetuini, D.; Cardone, D.; Chiarelli, A.M.; Merla, A. Thermal Infrared Imaging-Based Affective Computing and Its Application to Facilitate Human Robot Interaction: A Review. *Applied Sciences* **2020**, *10*. https://doi.org/10.3390/app10082924.
20. Kovács, K.E.; Őrsi, B.; Csukonyi, C.; Neamah, H.A.; Papp, D.; Korondi, P. A Systematic Review of Emotion Recognition and Non-Verbal Communication in Human-Robot Interaction. In Proceedings of the 2025 IEEE 16th International Conference on Cognitive Infocommunications (CogInfoCom), 2025, pp. 000183–000190. https://doi.org/10.1109/CogInfoCom66819.2025.11200706.
21. Zhao, M.; Gong, L.; Din, A.S. A review of the emotion recognition model of robots. *Applied Intelligence* **2025**, *55*, 364. https://doi.org/10.1007/s10489-025-06245-3.
22. Gasteiger, N.; Lim, J.; Hellou, M.; MacDonald, B.A.; Ahn, H.S. A Scoping Review of the Literature On Prosodic Elements Related to Emotional Speech in Human–Robot Interaction. *International Journal of Social Robotics* **2024**, *16*, 659–670. https://doi.org/10.1007/s12369-022-00913-x.
23. Cordeiro Ottoni, L.T.; de Jesus Fiais Cerqueira, J. A Review of Emotions in Human-Robot Interaction. In Proceedings of the 2021 Latin American Robotics Symposium (LARS), 2021 Brazilian Symposium on Robotics (SBR), and 2021 Workshop on Robotics in Education (WRE), 2021, pp. 7–12. https://doi.org/10.1109/LARS/SBR/WRE54079.2021.9605479.
24. Savery, R.; Weinberg, G. Robots and emotion: a survey of trends, classifications, and forms of interaction. *Advanced Robotics* **2021**, *35*, 1030–1042. https://doi.org/10.1080/01691864.2021.1957014.
25. Mitchell, J.J.; Jeon, M. Exploring Emotional Connections: A Systematic Literature Review of Attachment in Human-Robot Interaction. *International Journal of Human–Computer Interaction* **2025**, *41*, 11753–11774. https://doi.org/10.1080/10447318.2024.2445100.
26. Hieida, C.; Nagai, T. Survey and perspective on social emotions in robotics. *Advanced Robotics* **2022**, *36*, 17–32. https://doi.org/10.1080/01691864.2021.2012512.
27. Firmino de Souza, D.; Sousa, S.; Kristjuhan-Ling, K.; Dunajeva, O.; Roosileht, M.; Pentel, A.; Mõttus, M.; Can Özdemir, M.; Gratšjova, Ž. Trust and Trustworthiness from Human-Centered Perspective in Human–Robot Interaction (HRI)—A Systematic Literature Review. *Electronics*, *14*. https://doi.org/10.3390/electronics14081557.
28. Nilsson, N.J. *Principles of Artificial Intelligence*; Symbolic Computation, Springer-Verlag Berlin Heidelberg and Tioga Publishing Company: Berlin, Heidelberg, 1982. Jointly published with Tioga Publishing Company.
29. Russell, S.J.; Norvig, P. *Artificial Intelligence: A Modern Approach*, 3rd ed.; Prentice Hall, 2010.
30. Feil-Seifer, D.; Matarić, M.J. Human-Robot Interaction. In *Encyclopedia of Complexity and Systems Science*; Meyers, R.A., Ed.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2020; pp. 1–23. https://doi.org/10.1007/978-3-642-27737-5_274-5.
31. Mohamed, Y.; Ballardini, G.; Parreira, M.T.; Lemaignan, S.; Leite, I. Automatic Frustration Detection Using Thermal Imaging. In Proceedings of the 2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2022, pp. 451–459. https://doi.org/10.1109/HRI53351.2022.9889545.
32. Deng, J.; Pang, G.; Zhang, Z.; Pang, Z.; Yang, H.; Yang, G. cGAN Based Facial Expression Recognition for Human-Robot Interaction. *IEEE Access* **2019**, *7*, 9848–9859. https://doi.org/10.1109/ACCESS.2019.2891668.
33. Mamodiya, U.; Kishor, I.; Ahmed Syed, A.; Sankalkar, P.; Naik, N. An Adaptive Human–Robot Interaction Framework Using Real-Time Emotion Recognition and Context-Aware Task Planning. *IEEE Access* **2025**, *13*, 152219–152240. https://doi.org/10.1109/ACCESS.2025.3603738.
34. Huang, C.M.; Mutlu, B. Anticipatory robot control for efficient human-robot collaboration. In Proceedings of the 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2016, pp. 83–90. https://doi.org/10.1109/HRI.2016.7451737.

 doi:10.20944/preprints202603.2029.v1




35. Ryoo, M.S.; Fuchs, T.J.; Xia, L.; Aggarwal, J.K.; Matthies, L. Robot-Centric Activity Prediction from First-Person Videos: What Will They Do to Me? In Proceedings of the 2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2015, pp. 295–302.
36. Huggins, M.; Alghowinem, S.; Jeong, S.; Colon-Hernandez, P.; Breazeal, C.; Park, H.W. Practical Guidelines for Intent Recognition: BERT with Minimal Training Data Evaluated in Real-World HRI Application. In Proceedings of the 2021 16th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2021, pp. 341–350.
37. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); Burstein, J.; Doran, C.; Solorio, T., Eds., Minneapolis, Minnesota, 2019; pp. 4171–4186.
38. Li, N.; Ross, R. Hmm, You Seem Confused! Tracking Interlocutor Confusion for Situated Task-Oriented HRI. In Proceedings of the Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction, New York, NY, USA, 2023; HRI '23, p. 142–151. https://doi.org/10.1145/3568162.3576999.
39. Chakraborty, P.; Ahmed, S.; Yousuf, M.A.; Azad, A.; Alyami, S.A.; Moni, M.A. A Human-Robot Interaction System Calculating Visual Focus of Human's Attention Level. *IEEE Access* **2021**, *9*, 93409–93421. https://doi.org/10.1109/ACCESS.2021.3091642.
40. Celiktutan, O.; Skordos, E.; Gunes, H. Multimodal Human-Human-Robot Interactions (MHHRI) Dataset for Studying Personality and Engagement. *IEEE Transactions on Affective Computing* **2019**, *10*, 484–497. https://doi.org/10.1109/TAFFC.2017.2737019.
41. Salam, H.; Çeliktutan, O.; Hupont, I.; Gunes, H.; Chetouani, M. Fully Automatic Analysis of Engagement and Its Relationship to Personality in Human-Robot Interactions. *IEEE Access* **2017**, *5*, 705–721. https://doi.org/10.1109/ACCESS.2016.2614525.
42. Shen, Z.; Elibol, A.; Chong, N.Y. Inferring Human Personality Traits in Human-Robot Social Interaction. In Proceedings of the 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2019, pp. 578–579. https://doi.org/10.1109/HRI.2019.8673124.
43. Ramnauth, R.; Adéníran, E.; Adamson, T.; Lewkowicz, M.A.; Giridharan, R.; Reiner, C.; Scassellati, B. A Social Robot for Improving Interruptions Tolerance and Employability in Adults with ASD. In Proceedings of the 2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2022, pp. 4–13. https://doi.org/10.1109/HRI53351.2022.9889383.
44. Sandygulova, A.; Amirova, A.; Telisheva, Z.; Zhanatkyzy, A.; Rakhymbayeva, N. Individual Differences of Children with Autism in Robot-assisted Autism Therapy. In Proceedings of the 2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2022, pp. 43–52. https://doi.org/10.1109/HRI53351.2022.9889537.
45. Moharana, S.; Panduro, A.E.; Lee, H.R.; Riek, L.D. Robots for Joy, Robots for Sorrow: Community Based Robot Design for Dementia Caregivers. In Proceedings of the 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2019, pp. 458–467. https://doi.org/10.1109/HRI.2019.8673206.
46. Cruz-Sandoval, D.; Morales-Tellez, A.; Sandoval, E.B.; Favela, J. A Social Robot as Therapy Facilitator in Interventions to Deal with Dementia-related Behavioral Symptoms. In Proceedings of the 2020 15th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2020, pp. 161–169.
47. Hemminahaus, J.; Kopp, S. Towards Adaptive Social Behavior Generation for Assistive Robots Using Reinforcement Learning. In Proceedings of the 2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI, 2017, pp. 332–340.
48. Doğan, F.I.; Torre, I.; Leite, I. Asking Follow-Up Clarifications to Resolve Ambiguities in Human-Robot Conversation. In Proceedings of the 2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2022, pp. 461–469. https://doi.org/10.1109/HRI53351.2022.9889368.
49. Andriella, A.; Torras, C.; Alenyà, G. Short-Term Human–Robot Interaction Adaptability in Real-World Environments. *International Journal of Social Robotics* **2020**, *12*, 639–657. https://doi.org/10.1007/s12369-019-00606-y.
50. Chen, H.; Park, H.W.; Zhang, X.; Breazeal, C. Impact of Interaction Context on the Student Affect-Learning Relationship in Child-Robot Interaction. In Proceedings of the 2020 15th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2020, pp. 389–397.
51. Unhelkar, V.V.; Li, S.; Shah, J.A. Decision-Making for Bidirectional Communication in Sequential Human-Robot Collaborative Tasks. In Proceedings of the 2020 15th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2020, pp. 329–341.

 doi:10.20944/preprints202603.2029.v1




52. Devin, S.; Alami, R. An implemented theory of mind to improve human-robot shared plans execution. In Proceedings of the 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2016, pp. 319–326. https://doi.org/10.1109/HRI.2016.7451768.
53. Ramachandran, A.; Huang, C.M.; Scassellati, B. Give Me a Break! Personalized Timing Strategies to Promote Learning in Robot-Child Tutoring. In Proceedings of the 2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI, 2017, pp. 146–155.
54. Li, S.; Xu, L.; Yu, F.; Peng, K. Does Trait Loneliness Predict Rejection of Social Robots? The Role of Reduced Attributions of Unique Humanness : Exploring the Effects of Trait Loneliness on Anthropomorphism and Acceptance of Social Robots. In Proceedings of the 2020 15th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2020, pp. 271–280.
55. Ligthart, M.E.; Neerincx, M.A.; Hindriks, K.V. Memory-Based Personalization for Fostering a Long-Term Child-Robot Relationship. In Proceedings of the 2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2022, pp. 80–89. https://doi.org/10.1109/HRI53351.2022.9889446.
56. Görür, O.C.; Rosman, B.; Sivrikaya, F.; Albayrak, S. Social Cobots: Anticipatory Decision-Making for Collaborative Robots Incorporating Unexpected Human Behaviors. In Proceedings of the 2018 13th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2018, pp. 398–406.
57. Kwon, M.; Biyik, E.; Talati, A.; Bhasin, K.; Losey, D.P.; Sadigh, D. When Humans Aren't Optimal: Robots that Collaborate with Risk-Aware Humans. In Proceedings of the 2020 15th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2020, pp. 43–52.
58. Dino, F.; Zandie, R.; Abdollahi, H.; Schoeder, S.; Mahoor, M.H. Delivering Cognitive Behavioral Therapy Using A Conversational Social Robot. In Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2019, pp. 2089–2095. https://doi.org/10.1109/IROS40897.2019.8968576.
59. Kitt, E.R.; Crossman, M.K.; Matijczak, A.; Burns, G.B.; Kazdin, A.E. Evaluating the Role of a Socially Assistive Robot in Children's Mental Health Care. *Journal of Child and Family Studies* **2021**, *30*, 1722–1735. https://doi.org/10.1007/s10826-021-01977-5.
60. Laban, G.; Morrison, V.; Kappas, A.; Cross, E.S. Coping with Emotional Distress via Self-Disclosure to Robots: An Intervention with Caregivers. *International Journal of Social Robotics* **2025**, *17*, 1837–1870. https://doi.org/10.1007/s12369-024-01207-0.
61. Velner, E.; Boersma, P.P.; Graaf, M.M.d. Intonation in Robot Speech: Does it work the same as with people? In Proceedings of the 2020 15th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2020, pp. 569–578.
62. Hu, Y.; Hoffman, G. Using Skin Texture Change to Design Emotion Expression in Social Robots. In Proceedings of the 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2019, pp. 2–10. https://doi.org/10.1109/HRI.2019.8673012.
63. Pelikan, H.R.M.; Broth, M.; Keevallik, L. "Are You Sad, Cozmo?" How Humans Make Sense of a Home Robot's Emotion Displays. In Proceedings of the 2020 15th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2020, pp. 461–470.
64. Suguitan, M.; Gomez, R.; Hoffman, G. MoveAE: Modifying Affective Robot Movements Using Classifying Variational Autoencoders. In Proceedings of the 2020 15th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2020, pp. 481–489.
65. Block, A.E.; Christen, S.; Gassert, R.; Hilliges, O.; Kuchenbecker, K.J. The Six Hug Commandments: Design and Evaluation of a Human-Sized Hugging Robot with Visual and Haptic Perception. In Proceedings of the 2021 16th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2021, pp. 380–388.
66. Au, R.H.Y.; Ling, K.; Fraune, M.R.; Tsui, K.M. Robot Touch to Send Sympathy: Divergent Perspectives of Senders and Recipients. In Proceedings of the 2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2022, pp. 372–382. https://doi.org/10.1109/HRI53351.2022.9889419.
67. Dobrosovestnova, A.; Hannibal, G. Teachers' Disappointment: Theoretical Perspective on the Inclusion of Ambivalent Emotions in Human-Robot Interactions in Education. In Proceedings of the 2020 15th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2020, pp. 471–480.
68. Coyne, A.K.; Murtagh, A.; McGinn, C. Using the Geneva Emotion Wheel to Measure Perceived Affect in Human-Robot Interaction. In Proceedings of the 2020 15th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2020, pp. 491–498.

 doi:10.20944/preprints202603.2029.v1




69. Gordon, G.; Breazeal, C.; Engel, S. Can Children Catch Curiosity from a Social Robot? In Proceedings of the Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction, New York, NY, USA, 2015; HRI '15, p. 91–98. https://doi.org/10.1145/2696454.2696469.
70. Walker, N.; Weatherwax, K.; Allchin, J.; Takayama, L.; Cakmak, M. Human Perceptions of a Curious Robot that Performs Off-Task Actions. In Proceedings of the 2020 15th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2020, pp. 529–538.
71. Aliasghari, P.; Ghafurian, M.; Nehaniv, C.L.; Dautenhahn, K. Effects of Gaze and Arm Motion Kinesics on a Humanoid's Perceived Confidence, Eagerness to Learn, and Attention to the Task in a Teaching Scenario. In Proceedings of the Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction, New York, NY, USA, 2021; HRI '21, p. 197–206. https://doi.org/10.1145/3434073.3444651.
72. Briggs, G.; Chita-Tegmark, M.; Krause, E.; Bridewell, W.; Bello, P.; Scheutz, M. A Novel Architectural Method for Producing Dynamic Gaze Behavior in Human-Robot Interactions. In Proceedings of the 2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2022, pp. 383–392. https://doi.org/10.1109/HRI53351.2022.9889499.
73. Terzioğlu, Y.; Mutlu, B.; Şahin, E. Designing Social Cues for Collaborative Robots: The Role of Gaze and Breathing in Human-Robot Collaboration. In Proceedings of the 2020 15th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2020, pp. 343–357.
74. Pereira, A.; Oertel, C.; Fermoselle, L.; Mendelson, J.; Gustafson, J. Effects of Different Interaction Contexts when Evaluating Gaze Models in HRI. In Proceedings of the 2020 15th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2020, pp. 131–139.
75. Wit, J.d.; Brandse, A.; Krahmer, E.; Vogt, P. Varied Human-Like Gestures for Social Robots: Investigating the Effects on Children's Engagement and Language Learning. In Proceedings of the 2020 15th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2020, pp. 359–367.
76. Roy, L.; Croft, E.A.; Ramirez, A.; Kulić, D. GPT-Driven Gestures: Leveraging Large Language Models to Generate Expressive Robot Motion for Enhanced Human-Robot Interaction. *IEEE Robotics and Automation Letters* **2025**, *10*, 4172–4179. https://doi.org/10.1109/LRA.2025.3547631.
77. Song, H.; Zhang, Z.; Barakova, E.I.; Ham, J.; Markopoulos, P. Robot Role Design for Implementing Social Facilitation Theory in Musical Instruments Practicing. In Proceedings of the 2020 15th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2020, pp. 253–260.
78. Song, S.; Baba, J.; Nakanishi, J.; Yoshikawa, Y.; Ishiguro, H. Costume vs. Wizard of Oz vs. Telepresence: How Social Presence Forms of Tele-operated Robots Influence Customer Behavior. In Proceedings of the 2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2022, pp. 521–529. https://doi.org/10.1109/HRI53351.2022.9889665.
79. Kato, Y.; Kanda, T.; Ishiguro, H. May I help you? - Design of Human-like Polite Approaching Behavior-. In Proceedings of the 2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2015, pp. 35–42.
80. Jackson, R.B.; Williams, T.; Smith, N. Exploring the Role of Gender in Perceptions of Robotic Noncompliance. In Proceedings of the 2020 15th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2020, pp. 559–567.
81. Rea, D.J.; Schneider, S.; Kanda, T. "Is this all you can do? Harder!": The Effects of (Im)Polite Robot Encouragement on Exercise Effort. In Proceedings of the 2021 16th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2021, pp. 225–233.
82. Natarajan, M.; Gombolay, M. Effects of Anthropomorphism and Accountability on Trust in Human Robot Interaction. In Proceedings of the 2020 15th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2020, pp. 33–42.
83. Herse, S.; Vitale, J.; Johnston, B.; Williams, M.A. Using Trust to Determine User Decision Making & Task Outcome During a Human-Agent Collaborative Task. In Proceedings of the Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction, New York, NY, USA, 2021; HRI '21, p. 73–82. https://doi.org/10.1145/3434073.3444673.
84. van der Hoorn, D.P.; Neerincx, A.; de Graaf, M.M. "I think you are doing a bad job!" : The Effect of Blame Attribution by a Robot in Human-Robot Collaboration. In Proceedings of the 2021 16th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2021, pp. 140–148.
85. Bryant, D.; Borenstein, J.; Howard, A. Why Should We Gender? The Effect of Robot Gendering and Occupational Stereotypes on Human Trust and Perceived Competency. In Proceedings of the 2020 15th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2020, pp. 13–21.

 doi:10.20944/preprints202603.2029.v1




86. Torre, I.; Linard, A.; Steen, A.; Tumová, J.; Leite, I. Should Robots Chicken? How Anthropomorphism and Perceived Autonomy Influence Trajectories in a Game-Theoretic Problem. In Proceedings of the 2021 16th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2021, pp. 370–379.
87. Hover, Q.R.; Velner, E.; Beelen, T.; Boon, M.; Truong, K.P. Uncanny,Sexy,and Threatening Robots : The Online Community's Attitude to and Perceptions of Robots Varying in Humanlikeness and Gender. In Proceedings of the 2021 16th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2021, pp. 119–128.
88. Iizawa, D.; Yamanaka, S. Face on a Globe: A Spherical Robot that Appears Lifelike Through Smooth Deformations and Autonomous Movement. In Proceedings of the 2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2022, pp. 502–510. https://doi.org/10.1109/HRI53351.2022.9889453.
89. Löffler, D.; Dörrenbächer, J.; Hassenzahl, M. The Uncanny Valley Effect in Zoomorphic Robots: The U-Shaped Relation Between Animal Likeness and Likeability. In Proceedings of the 2020 15th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2020, pp. 261–270.
90. Bajcsy, A.; Losey, D.P.; O'Malley, M.K.; Dragan, A.D. Learning from Physical Human Corrections, One Feature at a Time. In Proceedings of the 2018 13th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2018, pp. 141–149.
91. Losey, D.P.; Bajcsy, A.; O'Malley, M.K.; Dragan, A.D. Physical interaction as communication: Learning robot objectives online from human corrections. *The International Journal of Robotics Research* **2022**, *41*, 20–44. https://doi.org/10.1177/02783649211050958.
92. Basu, C.; Singhal, M.; Dragan, A.D. Learning from Richer Human Guidance: Augmenting Comparison-Based Learning with Feature Queries. In Proceedings of the 2018 13th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2018, pp. 132–140.
93. Chen, L.; Paleja, R.; Ghuy, M.; Gombolay, M. Joint Goal and Strategy Inference across Heterogeneous Demonstrators via Reward Network Distillation. In Proceedings of the 2020 15th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2020, pp. 659–668.
94. Bobu, A.; Scobee, D.R.R.; Fisac, J.F.; Sastry, S.S.; Dragan, A.D. LESS is More: Rethinking Probabilistic Models of Human Behavior. In Proceedings of the Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction, New York, NY, USA, 2020; HRI '20, p. 429–437. https://doi.org/10.1145/3319502.3374811.
95. McQuillin, E.; Churamani, N.; Gunes, H. Learning Socially Appropriate Robo-waiter Behaviours through Real-time User Feedback. In Proceedings of the 2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2022, pp. 541–550. https://doi.org/10.1109/HRI53351.2022.9889395.
96. Moreira, I.; Rivas, J.; Cruz, F.; Dazeley, R.; Ayala, A.; Fernandes, B. Deep Reinforcement Learning with Interactive Feedback in a Human–Robot Environment. *Applied Sciences* **2020**, *10*. https://doi.org/10.3390/app10165574.
97. Celemin, C.; Kober, J. Knowledge- and ambiguity-aware robot learning from corrective and evaluative feedback. *Neural Computing and Applications* **2023**, *35*, 16821–16839. https://doi.org/10.1007/s00521-022-08118-z.
98. Chisari, E.; Welschehold, T.; Boedecker, J.; Burgard, W.; Valada, A. Correct Me If I am Wrong: Interactive Learning for Robotic Manipulation. *IEEE Robotics and Automation Letters* **2022**, *7*, 3695–3702. https://doi.org/10.1109/LRA.2022.3145516.
99. van Waveren, S.; Pek, C.; Tumova, J.; Leite, I. Correct Me If I'm Wrong: Using Non-Experts to Repair Reinforcement Learning Policies. In Proceedings of the 2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2022, pp. 493–501. https://doi.org/10.1109/HRI53351.2022.9889604.
100. Schrum, M.L.; Hedlund-Botti, E.; Moorman, N.; Gombolay, M.C. MIND MELD: Personalized Meta-Learning for Robot-Centric Imitation Learning. In Proceedings of the 2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2022, pp. 157–165. https://doi.org/10.1109/HRI53351.2022.9889616.
101. Trung, P.; Giuliani, M.; Miksch, M.; Stollnberger, G.; Stadler, S.; Mirnig, N.; Tscheligi, M. Head and shoulders: automatic error detection in human-robot interaction. In Proceedings of the Proceedings of the 19th ACM International Conference on Multimodal Interaction, New York, NY, USA, 2017; ICMI '17, p. 181–188. https://doi.org/10.1145/3136755.3136785.
102. Kontogiorgos, D.; Pereira, A.; Sahindal, B.; Waveren, S.v.; Gustafson, J. Behavioural Responses to Robot Conversational Failures. In Proceedings of the 2020 15th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2020, pp. 53–62.

 doi:10.20944/preprints202603.2029.v1




103. Loureiro, F.; Avelino, J.; Moreno, P.; Bernardino, A. Self-perception of Interaction Errors Through Human Non-verbal Feedback and Robot Context. In Proceedings of the Social Robotics; Cavallo, F.; Cabibihan, J.J.; Fiorini, L.; Sorrentino, A.; He, H.; Liu, X.; Matsumoto, Y.; Ge, S.S., Eds., Cham, 2022; pp. 475–487.
104. Kwon, M.; Huang, S.H.; Dragan, A.D. Expressing Robot Incapability. In Proceedings of the 2018 13th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2018, pp. 87–95.
105. Stange, S.; Kopp, S. Effects of a Social Robot's Self-Explanations on How Humans Understand and Evaluate Its Behavior. In Proceedings of the 2020 15th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2020, pp. 619–627.
106. Das, D.; Banerjee, S.; Chernova, S. Explainable AI for Robot Failures: Generating Explanations that Improve User Assistance in Fault Recovery. In Proceedings of the 2021 16th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2021, pp. 351–360.
107. Green, H.N.; Islam, M.M.; Ali, S.; Iqbal, T. Who's Laughing NAO? Examining Perceptions of Failure in a Humorous Robot Partner. In Proceedings of the 2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2022, pp. 313–322. https://doi.org/10.1109/HRI53351.2022.9889353.
108. Tang, B.; Sullivan, D.; Cagiltay, B.; Chandrasekaran, V.; Fawaz, K.; Mutlu, B. CONFIDANT: A Privacy Controller for Social Robots. In Proceedings of the Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction. IEEE Press, 2022, HRI '22, p. 205–214.
109. Dorafshanian, M.; Aitsam, M.; Mejri, M.; Di Nuovo, A. Beyond Data Collection: Safeguarding User Privacy in Social Robotics. In Proceedings of the 2024 IEEE International Conference on Industrial Technology (ICIT), 2024, pp. 1–6. https://doi.org/10.1109/ICIT58233.2024.10540743.
110. Aryania, A.; Chockalingam, S.; Rødsethol, H.K.; Alenya, G. Impact of Design Transparency on Trust and Data Sharing During Human-Robot Interactions in Public Places. *J. Hum.-Robot Interact.* **2025**. Just Accepted, https://doi.org/10.1145/3785152.
111. Jackson, R.B.; Williams, T. Language-Capable Robots may Inadvertently Weaken Human Moral Norms. In Proceedings of the 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2019, pp. 401–410. https://doi.org/10.1109/HRI.2019.8673123.
112. Wen, R.; Han, Z.; Williams, T. Teacher, Teammate, Subordinate, Friend: Generating Norm Violation Responses Grounded in Role-based Relational Norms. In Proceedings of the 2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2022, pp. 353–362. https://doi.org/10.1109/HRI53351.2022.9889594.
113. Lacey, C.; Caudwell, C. Cuteness as a 'Dark Pattern' in Home Robots. In Proceedings of the 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2019, pp. 374–381. https://doi.org/10.1109/HRI.2019.8673274.
114. Winkle, K.; Caleb-Solly, P.; Leonards, U.; Turton, A.; Bremner, P. Assessing and Addressing Ethical Risk from Anthropomorphism and Deception in Socially Assistive Robots. In Proceedings of the 2021 16th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2021, pp. 101–109.
115. Sharkey, A.; Sharkey, N. We need to talk about deception in social robotics! *Ethics and Information Technology* **2021**, *23*, 309–316. https://doi.org/10.1007/s10676-020-09573-9.
116. Winkle, K.; Jackson, R.B.; Melsión, G.I.; Brščić, D.; Leite, I.; Williams, T. Norm-Breaking Responses to Sexist Abuse: A Cross-Cultural Human Robot Interaction Study. In Proceedings of the 2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2022, pp. 120–129. https://doi.org/10.1109/HRI53351.2022.9889389.
117. Trainer, T.; Taylor, J.R.; Stanton, C.J. Choosing the Best Robot for the Job: Affinity Bias in Human-Robot Interaction. In Proceedings of the Social Robotics, Cham, 2020; pp. 490–501.
118. Barfield, J. Designing Social Robots to Accommodate Diversity, Equity, and Inclusion in Human-Robot Interaction, New York, NY, USA, 2023; CHIIR '23, p. 463–466. https://doi.org/10.1145/3576840.3578303.
119. Londoño, L.; Valeria Hurtado, J.; Hertz, N.; Kellmeyer, P.; Voeneky, S.; Valada, A. Fairness and Bias in Robot Learning. *Proceedings of the IEEE* **2024**, *112*, 305–330. https://doi.org/10.1109/JPROC.2024.3403898.
120. Fossa, F.; Sucameli, I. Gender Bias and Conversational Agents: an ethical perspective on Social Robotics. *Science and Engineering Ethics* **2022**, *28*, 23. https://doi.org/10.1007/s11948-022-00376-3.

 doi:10.20944/preprints202603.2029.v1



## Short Biography of Authors

**Yunjia Sun** is currently a postdoctoral researcher at School of Computer Science, Peking University. She received the BS and PhD degrees from University of Chinese Academy of Sciences, Beijing, China, in 2018 and 2024, respectively. Her research interests include computer vision, affective computing, and human-robot interaction.Her research has been published in top-tier venues, including leading conferences and journals such as ICCV and TPAMI.

**Tao Wang** is a Research Professor in School of Computer Science in Peking University, and the director of the Laboratory for Affective and Cognitive Intelligent Robotics. He received B.S. and Ph.D. degrees from Peking University in 1999 and 2006, respectively. He has published more than 80 research papers, many of which were in top-conferences such as AAAI, MULTIMEDIA, IROS, IJCAI, ISCA, MICRO, HPCA, MobiCom, and MobiSys, and in premier journals including IEEE TC, IEEE TMC, IEEE TWC and IEEE TCAD. He has been authorized more than 20 invention patents. He won Best Community Paper Award in MobiCom'17 (contributes the most to the broader research community), and some Best Paper Awards in other conferences. He won the award of "2008 Intel China Employee of the Year," the highest individual award at Intel China. The textbook *Embodied Intelligent Robotics: Principles and Practices* edited by him has been selected into China's "14th Five-Year Plan" National Key Publication Planning. He once served as a Qianjiang Distinguished Expert in Hangzhou, China. He is currently serving as the Secretary-General of Beijing Computer Federation. He also serves as an editorial board member of International Journal of Crowd Science, and an editorial board member of Electronics (Artificial Intelligence section). His current research interests are: computer achitecture, affective and cognitive computing, human-robot interaction, and intelligent robot system.