A Taxonomy of Social Errors in Human-Robot Interaction

LEIMIN TIAN, Faculty of Engineering, Monash University, Australia
SHARON OVIATT, Faculty of Engineering, Monash University, Australia

Robotic applications have entered various aspects of our lives, such as health care and educational services. In such Human-Robot Interaction (HRI), trust and mutual adaption is established and maintained through a positive social relationship between a user and a robot. This social relationship relies on the perceived competence of a robot on the social-emotional dimension. However, because of technical limitations and user heterogeneity, current HRI is far from error-free, especially when a system leaves controlled lab environments and is applied to in-the-wild conditions. Errors in HRI may either degrade a user's perception of a robot's capability in achieving a task (defined as **performance errors** in this work), or degrade a user's perception of a robot's socio-affective competence (defined as **social errors** in this work). The impact of these errors and effective strategies to handle such impact remains an open question. We focus on social errors in HRI in this work. In particular, we identify major attributes of perceived socio-affective competence by reviewing human social interaction studies and HRI error studies. This motivates us to propose a taxonomy of social errors in HRI. We then discuss the impact of social errors situated in three representative HRI scenarios. This paper provides foundations for a systematic analysis of the social-emotional dimension of HRI. The proposed taxonomy of social errors encourages the development of user-centered HRI systems, which offer positive and adaptive interaction experiences and improved interaction outcomes.

CCS Concepts: • **Human-centered computing** → **Collaborative and social computing**; **Collaborative and social computing theory, concepts and paradigms**; • **Applied computing** → *Sociology*.

Additional Key Words and Phrases: Affective computing, human-robot interaction, social robotics, social norms, socio-affective competence

ACM Reference Format:
Leimin Tian and Sharon Oviatt. 2020. A Taxonomy of Social Errors in Human-Robot Interaction. *ACM Trans. Hum.-Robot Interact.* 0, 0, Article 0 (2020), 34 pages. https://doi.org/0000001.0000001

1 INTRODUCTION

David checks into the Henn-na (meaning "Strange") Hotel in Japan, a place staffed by robots to give its visitors a taste of the future. He unlocks his room, and on the nightstand sits a pink doll-shaped robot named Churi with a friendly smile painted on its face. "Relax and make yourself at home." It speaks in a dry robotic voice. "Hello?" David asks with a bit of hesitation in his voice. "Oh. Hi, there. How do you do?" The robot replies which puts a smile on David's face. "Really good." David answers, and continues: "Can you sing a song?" There is no response. David adds: "Just whatev..." Before he can finish, the robot interrupts:

Authors' addresses: Leimin Tian, Faculty of Engineering, Monash University, 14 Alliance Lane, Clayton, VIC, 3800, Australia, Leimin.Tian@monash.edu; Sharon Oviatt, Faculty of Engineering, Monash University, 14 Alliance Lane, Clayton, VIC, 3800, Australia, Sharon.Oviatt@monash.edu.

2573-9522/2020/0-ART0 $15.00
https://doi.org/0000001.0000001

"What time are you going to wake up?" "Eh..." Now the smile on his face has disappeared. David shakes his head and sighs. "Seven a.m. But back to the original question, can you sing a song?" The robot answers: "Yes. May I help you?" David leans foward and repeats his question slowly: "Yes. CAN YOU SING A SONG?!" The robot replies: "Tomorrow's set an alarm clock at eight-thirty." David looks away and frowns.[1] What went wrong here? Beyond technical failures such as speech recognition errors, what happened in addition that caused the unease David felt?

With recent technical advances, we have witnessed significant growth in the application of Artificial Intelligence and robotics in various domains of our society (see Sheridan [162] for a review). Because of the importance of emotions in human cognition and communication, it is inevitable for researchers to take emotions into account, which led to the establishment of the field Affective Computing [136]. In current Human-Robot Interaction (HRI) research, the social-emotional dimension of interaction has drawn an increasing attention. An HRI system able to recognize and react to emotional states of its users has various potential benefits. For example, in elderly care applications, HRI designs that incorporate social-emotional interactions have been shown to improve health outcomes, by encouraging positive moods and reducing loneliness felt by users [58].

Current social-emotional interaction functions of HRI systems are far below human-level performance, and errors are inevitable during interactions. These errors may have a significant impact on a user's perception of a robot. However, error handling in HRI is not yet fully understood, especially regarding errors on the social-emotional dimension. Previous research on errors in HRI has focused on errors in the functional requirements of a robot, such as navigation [156]. However, compliance and adaption to mutually agreed social norms are essential to establishing social relationships [170]. Thus, when a robot violates social norms, it may cause significant influence on a user's perceptions, and understanding such influence may be a key to advance current HRI research, especially in longitudinal scenarios where a positive human-robot relationship is desired.

Errors are also opportunities for improvement. Levine et al. [100] found that more reliable and effective hand-eye coordination for robotic grasping can be learned through correcting errors. By addressing errors on the social-emotional dimension of HRI and identifying their impact, we can potentially develop more personalized, adaptive, and socially acceptable HRI. This is crucial for achieving improved outcomes of HRI applications, such as better study outcomes in educational applications, more humane health-care, and more effective human-robot collaborative problem solving.

Error identification and recovery remains a challenging topic in current HRI research, especially on the social-emotional dimension. This motivates us to formally define social errors in HRI and provide a systematic analysis of the key attributes and impact of such errors. In particular, we will introduce a taxonomy of social errors in HRI from a user-centered perspective, which we expect will serve as a foundation for advancing socio-affective competence of current HRI systems.

1.1 Social errors and performance errors in HRI

We consider an "error" as a behavior exhibited by a robot that deviates from the desired or normal behaviors expected by a user. This definition covers two types of exceptions in HRI, namely technical exceptions in delivering a designed function, and perceived exceptions

[1] Adapted from episode 2 "Japan" of David Farrier's 2018 documentary series *Dark Tourist* on Netflix: https://www.netflix.com/au/title/80189791

in which a robot delivers a function as designed, but its behavior is not the ideal to an individual user. Our discussion of HRI errors will focus on the *perceived* exceptions. We use the term “breach” interchangeably with “error” and “failure” in this manuscript, because they all refer to such deviations from user’s expectations.

Existing research on errors in HRI investigates errors either from a causal perspective, such as classifying them into physical errors and human errors [23], or from the consequence perspective, such as classifying them into benign errors and catastrophic errors [94]. In our work, we analyze errors in HRI from a user-centered design perspective. User-centered design focuses on addressing the needs and preferences of users and the intended application scenarios of the designed systems [124]. It is commonly adopted in human-computer interaction design practices, and has known benefits such as improved usability and user satisfaction [93]. Following the user-centered perspective, we classify errors in HRI as performance errors and social errors with the operational definitions below based on their influence on a user’s perceptions:

- **Performance errors** are errors that degrade a user’s perception of a robot’s intelligence and competence in achieving a task, such as failure to register a spoken command in a noisy environment; and
- **Social errors** are errors that violate social norms and degrade a user’s perception of a robot’s socio-affective competence and their relationship with it, such as interrupting a user at an inappropriate time during a conversation.

A social error may be caused by technical malfunctions, such as delayed dialogue responses, or by imperfect design of social-emotional interaction functions of a HRI system, such as not incorporating individual variances. An error scenario can be a mix of performance error and social error. For example, a robot not able to register a spoken command repeatedly may irritate a user and result in abortion of the interaction. In this paper, we focus on understanding these errors and their *impact*, rather than their causes. Our discussion of various impact of social errors in HRI will be situated in context, such as individual differences in users or goals of an HRI system.

1.2 Research question and hypotheses

The main research question we address in this paper is how to systematically analyze social errors in HRI. Our hypothesis is that we can develop a taxonomy of social errors applicable to HRI based on psychology and social cognition theories of human social interactions and interpersonal communications. We expect the proposed taxonomy to guide a structured analysis of the impact of social errors in HRI on the perceived *socio-affective competence* of a robot and the *social relationship* between a robot and its users.

Various definitions of socio-affective competence and social relationship exist in the literature. Here we adopt the following definitions:

- **Socio-affective competence** is the ability to successfully conduct social interactions, which depends on the awareness and identification of social-emotional cues, the ability to process such cues, and the ability to decide on and express a normative response to these cues [62];
- **Social relationship** is a connection between a person and another entity, which represents the person’s perception of the availability or adequacy of resources and assistance provided by that other entity, and results in interdependency of their behaviors [26].

1.3 Road map

The remainder of this paper is organized as follows: In Section 2, we review studies on human emotions and social interaction to identify major attributes of socio-affective competence. In Section 3, we review existing HRI studies on errors and their influences. We then propose a taxonomy of social errors in HRI in Section 4 supported by our multidisciplinary review. We discuss each type of social error in the proposed taxonomy in representative HRI scenarios and existing HRI studies. Finally, in Section 5, we summarize our major contributions and provide future directions of our work.

2 SOCIO-AFFECTIVE COMPETENCE IN INTERPERSONAL COMMUNICATIONS

In this section, we review socio-affective competence and social relationships in human-human interaction. This allows us to decompose the impact of HRI social errors into individual dimensions that people perceive socio-affective competence on, thus informing the categories of HRI social errors in our proposed taxonomy. We start with an overview of emotion and its function in Section 2.1 due to the central role emotion plays in social communication. We continue to discuss empathy and entrainment in Section 2.2 as they form the basis of social relationships. To extend the discussion beyond healthy, general population, we proceed to analyze assessment of impaired social-emotional functions caused by mental health disorders in Section 2.3. Finally, we zoom out to the organizational and societal level and review social norms in the context of prosociality in Section 2.4. Section 2.5 summarizes our discussions on main attributes of socio-affective competence and how they relate to the analysis of HRI social errors. During the discussion in each section, we connect the socio-affective competence attributes identified from human studies to existing HRI studies, which provides supports to the taxonomy of social errors in HRI proposed in this work. Furthermore, we summarize common assessment approaches in human studies when evaluating each attribute of socio-affective competence: For objective or behavior-based assessments, researchers may consider replicating them to evaluate the impact of HRI social errors on those perceived socio-affective competence dimensions; For socio-affective competence attributes relying on subjective or self-report style assessments, they call for HRI researchers to make efforts in developing novel assessment approaches in the HRI context.

2.1 Emotion and social interaction

Psychology and neuroscience studies identified three main *intra-personal* functions of emotions: (1) the attention-directing function, i.e., emotions allow better allocation of limited cognitive resources [163]; (2) the informational function, i.e., emotions provide high-level evaluation of events based on their relationship with someone's internal beliefs and desires [165]; and (3) the motivational function, i.e., emotions serve as a reward function which motivates us to strategize our behaviors based on emotional feedback [126]. These functions of emotion also motivated researchers to implement artificial emotions in a robot to improve its ability to generate adaptive behaviors [1, 117].

During *inter-personal* communications, emotions have been identified as cues that allow people to observe and perceive the latent cognitive states of their interaction partners, and to express their own internal states [153, 176]. Emotional expressions and reactions to other's emotions were found to be essential for establishing social relationships [103]. Emotional experiences navigate us through the complexities of the social world [46]. For example, while monitoring our progress towards achieving a specific social goal (e.g., the desire to affiliate with someone), emotional experiences can act as a reward function which motivates us to adjust our decision policies and action strategies during social interaction. Thus, for a robot to achieve social goals, it also requires emotional interaction functions, including identifying, processing, and expressing emotions during inter-personal communication.

There have been evidences that emotional experiences and expressions are personal and context specific, i.e., "variation is the norm" [11]. The same event or stimulus can evoke a spectrum of emotional responses in different people and can be perceived differently, as demonstrated in existing studies of emotion induction and annotation tasks [137]. Behaviors and physiological responses across individuals also differ when they experience the same emotion. For example, a previous study identified that people can have varied actions in response to fear, such as freeze or flee, with different neural circuits involved [60]. Emotions are also cultural-specific in their definition and expression. For instance, some languages do not make clear distinction between anger and sadness [151], while some emotional lexicons are "untranslatable" across languages [127]. In addition, a previous cross-cultural study found that norms in smiling, laughter, and expressions of positive emotions were related to the historical heterogeneity of the population in a country or a state [122]. Therefore, context and individual variances need to be taken into account for enabling personalized and adaptive HRI experiences. Emotions are also temporally dependent. They relate to long-term factors, such as moods or personality traits [110], as well as other cognitive processes, such as memory or attention [135]. Thus, emotional interaction functions of an HRI system need to be integrated with other system modules, for instance, its dialogue manager.

Our review of the functions of human emotions demonstrated that emotions are fundamental to intelligence and interaction. The attributes of socio-affective competence related to emotional interaction are summarized in Table 1. For a robot to facilitate social interaction with its users, it requires the ability to recognize emotions of its users, to incorporate in it's decision-making process the emotions of users and artificial emotions generated by itself, and to synthesize emotional expressions according to the interaction context. Failures in doing so will result in HRI social errors on these socio-affective competence dimensions, which will be discussed in more details in Section 4.

2.2 Empathy and social relationship

As reviewed in the previous section, emotion is an important component of interpersonal communications. One particular emotional phenomenon that attracts growing attention in recent HRI studies is the computational analysis and simulation of empathy [181]. Cuff et al. [29] summarized empathy as an emotional response that "is similar to one's perception and understanding of the stimulus emotion, with recognition that the source of the emotion is not one's own". Decety and Meyer [35] also identified that the shared anchor of different definitions of empathy is the matching relation between the emotions of two people. In this section, we review the functions of empathy and assessments of a person's ability to exhibit empathy in social interaction. This allows us to identify socio-affective competence attributes related to empathy and social relationships, which provides support to HRI researchers developing and evaluating an empathic robot.

Empathy allows the development of shared representations, engagement, and simulation of the subjectivity of others. It is also important for the formation and maintenance of social relationships [34]. It often leads to behavioral and relational outcomes. For example, in psychotherapy sessions, empathy of a therapist is a key behavioral markers for the quality and constructive outcomes of the therapy [21]. Therefore, an HRI system with empathic functions leads to better understanding of the users, and is required for establishing social relationships between the system and its users.

Empathic and emotional reactions	
Socio-affective competence attributes	**Assessment methods**
Ability to recognize and understand emotions from various cues	Objective, e.g., [57]
Ability to express emotions through verbal and non-verbal communicative modalities	Subjective, e.g., [144]
Ability to self-regulate and stabilize emotions	Subjective, e.g., [111]
Responsiveness and synchronicity to other's emotions	Objective, e.g., [68]
Ability to exhibit entrainment responses to promote mutual understanding	Objective, e.g., [88]
Ability of perspective-taking to support understanding of others	Subjective, e.g., [31]
Social skills and social relationship	
Socio-affective competence attributes	**Assessment methods**
Ability to recognize and understand social relationships and social contexts	Objective, e.g., [9]
Ability to adjust behaviors according to appropriate intimacy level of social relationships and social contexts	Objective, e.g., [66]
Ability to maintain independence and exhibit assertiveness	Subjective, e.g., [106]
Awareness and reasonable adherence to social norms to facilitate goals	Subjective, e.g., [138]
Ability to engage in reciprocal social communication	Objective, e.g., [105]
Anticipation of social routine	Objective, e.g., [105]
Ability to establish and maintain social relationships	Subjective, e.g., [71]

Table 1. Socio-affective competence attributes related to emotional interaction and social skills.

Other than empathy, the matching of emotion between people has also been studied in emotional contagion. Emotional contagion is "the tendency to automatically mimic and synchronize expressions, vocalizations, postures, and movements with those of another person's and, consequently, to converge emotionally" [65]. Compared to empathy, emotional contagion focuses on convergence of verbal and non-verbal behaviors in interpersonal communications, or entrainment [120]. Co-occurrence in behavioral patterns of emotional expressions between conversational dyads has been shown to be a predictor of rapport in collaborative problem solving [108]. The degree of convergence in human dyadic interaction has also been found to be positively correlated with the overall success of collaboration and positive attitudes of team members [101]. In educational scenarios, entrainment has been identified as an indicator for better learning outcomes and mutual understanding [128]. These indicate that entrainment can be utilized in HRI design to increase the engagement of users, as well as to improve interaction outcomes, such as users' learning outcomes of an educational robot.

The ability to exhibit empathy and entrainment is an important factor evaluated in various assessments of social skills, such as the Social Skills Inventory [106], the Mayer-Salovey-Caruso Emotional Intelligence Test [111], and the Self Descriptive Inventory [144]. Commonly used empathy assessments include the Empathy Quotient [10], the Empathy Scale [76], the Interpersonal Reactivity Index [31], and the Questionnaire Measure of Emotional Empathy [113]. These assessments of empathy and social skills share attributes that reflect the main dimensions of human's perception of empathy and social skills of self and others. We summarize these shared socio-affective competence dimensions and their assessment methods in Table 1.

When designing and evaluating robots capable of emotional interactions, HRI researchers may refer to the dimensions of socio-affective competence in Table 1, which arise from our review of emotions, empathy, and social skills. These attributes can be used to measure users' perception of social errors and their impacts in such HRI systems. For example, Trinh et al. [172] developed a robotic coach for providing feedbacks on people's oral presentation skills. The goal of this HRI system is to help its user improve their presentation quality, while reducing their anxiety and increasing their confidence. The robot was designed to give positive feedback with verbal emotional expressions, such as "great job", added to suggestions to a user's presentation performance. The identified dimensions of socio-affective competence in Table 1 suggest that this HRI system may be further improved, for instance, by synthesizing emotional expressions responsive to a user's emotions, or by exhibiting assertiveness in its feedback on a user's presentation performance.

2.3 Socio-affective competence and mental health

Our review so far focused on socio-affective competence of a healthy, general population. However, one important application scenario of HRI is mental health care. In particular, existing HRI studies have focused on two domains [143]: 1) interventions for individuals with impaired cognitive and social abilities, and 2) providing companionship as well as physical and mental supports to individuals living on their own. A representative of interventions for individuals with impaired mental abilities is the use of social robots for diagnosis and training of children with Autism Spectrum Disorders (ASD) [39]. Regarding providing companionship and care to individuals living alone, a representative is the use of social robots for elderly care, especially for people with impaired cognitive functions, such as dementia patients [119]. HRI applications in mental health extend beyond ASD and elderly care, e.g., HRI for depression care [15]. However, we focus on these two mental health domains as representatives, and summarize major assessments for impairment and atypicality in a person's social communicative functions. Our review provides a user-centered perspective for identifying HRI social errors in mental health applications, and for evaluating the effectiveness of robot-assisted treatment.

2.3.1 Autism spectrum disorders.

Autism Spectrum Disorder (ASD) refers to a class of neuro-developmental disorders [51]. A major diagnostic cue of ASD is impairment in reciprocal social interactions and communications [105]. We review existing psychometric assessments for ASD diagnosis and treatment, which allows us to summarize attributes of socio-affective competence relating to reciprocal social interactions and communications.

A commonly used assessment of atypical social communicative functions is evaluating a person's ability to understand others. For example, asking an individual to describe emotions expressed in photos of faces [8] or speech recordings [152]. To capture the multimodal and context-specific nature of social interaction, assessments with movie clips were also used, such as the Movie for the Assessment of Social Cognition task [43]. Beyond emotion recognition tests, self-reported evaluations of social interaction capabilities have also been used, for example, the empathy quotient questionnaire [10]. However, such tests may not be applicable to those with limited verbal ability, or across different languages or cultures.

For individuals with limited communicative ability, an observer's evaluation can be used to assess their development level. A widely used assessment following such an approach is the Autism Diagnostic Observation Schedule [59]. Representative behavior cues evaluated by an observer include a person's use of verbal or non-verbal expressions, response time in conversations, use of eye contact or gaze, amount and quality of reciprocal social communication, anticipation of social routines, empathic or emotional reactions, and shared enjoyment and joint attention.

2.3.2 Age-related decline in social and cognitive abilities.

Decline in cognitive functions occurs in elderly adults due to the loss of neurons and neuronal connections in the brain. The progressive impairment can lead to reduced memory function, reduced ability in understanding other's thoughts and emotions, altered communication, depressed and unstable mood, and feelings of isolation and loneliness [91]. In healthy elderly adults, impairment in social and cognitive functions may be mild and does not necessarily result in significant decline in the quality of their personal and social life [158]. Current research on healthy aging has focused on identifying reliable methods to improve the emotional well-being and social functioning of individuals, and preventing further decline of cognitive functions [158]. Elderly adults with neurodegenerative diseases typically experience more severe impairment in social cognition functions, which significantly impacts their ability to maintain independence and engage in social interactions. For example, individuals with frontotemporal dementia [146] were found to have difficulties recognizing facial expressions and gaze directions.

Similar to assessments of social interaction capability used in ASD, emotion recognition and social context interpretation tasks have been applied to assess the degree of age-related social cognition decline, such as the Edinburgh social cognition test [7]. Questionnaires answered by an elderly person or their carers have also been used. In particular, three standardized evaluation questionnaires are widely used for dementia patients or people at risk, namely the interpersonal reactivity index [30], the social norms questionnaire [138], and the social behavior observer checklist [139]. The interpersonal reactivity index measures empathic concerns and ability of perspective taking. The social norms questionnaire requires an individual to assess if a behavior is appropriate or not in order to evaluate whether the person is overly adherent to a perceived social norm or unaware of social norm violations. The social behavior observer checklist collects a carer's observations on decreased social functions of an individual, such as showing diminished social and emotional engagement, having unstable emotional and cognitive abilities, exhibiting perseveration, or not being sensitive to others' embarrassment or privacy.

2.3.3 Socio-affective competence in mental health contexts.

Shared attributes of socio-affective competence in ASD assessments and cognitive impairment assessments are reported in Table 2. Note that we omitted from Table 2 the attributes that are already included in Table 1, such as the emotion recognition attribute. These attributes from mental health context will help us extend the dimensions that HRI social errors can be identified and measured on. Furthermore, these identified attributes of socio-affective competence will inform an user-centered design and evaluation of HRI systems for assistive and mental health applications. For example, Paletta et al. [132] used a Pepper robot to motivate dementia patients to interact with therapeutic games installed on the robot's chest tablet. These games offer exercises for the improvement of auditive and visual working memory. The robot estimates a user's emotional states and exercise scores, and lowers the difficulty of the games when the user shows signs of disengagement. The socio-affective competence attributes in Table 2 suggests that this HRI system may be further improved. For example, instead of lowering the game difficulty, which limits the intensity of exercise a user receives, the robot may motivate the user by displaying shared enjoyment that celebrates game successes using verbal and non-verbal behaviors.

Understanding others	
Socio-affective competence attributes	**Assessment methods**
Ability to understand knowledge-state of others	Subjective, e.g., [9]
Ability to understand intentions of others	Objective, e.g., [43]
Responsiveness to joint attention and eye gaze	Objective, e.g., [59]
Responsiveness to shared enjoyment	Objective, e.g., [59]
Self-consciousness and anticipation of others' reception of self	Subjective, e.g., [30]
Communicative functions	
Socio-affective competence attributes	**Assessment methods**
Ability to initiate verbal and non-verbal communications	Objective, e.g., [59]
Ability to respond to verbal and non-verbal communications with relevant and comprehensible replies	Objective, e.g., [59]
Ability to use appropriate non-verbal expressions to accommodate communication	Objective, e.g., [59]
Ability to perform turn-taking at appropriate times	Objective, e.g., [59]
Ability to maintain engagement of self and others	Subjective, e.g., [30]
Ability to maintain coherence and adaption during communication	Subjective, e.g., [30]

Table 2. Socio-affective competence attributes related to reciprocal social communication.

2.4 Social norms as studied in behavioral economics

In previous sections, our review has focused on identifying main attributes of socio-affective competence in interpersonal communications. Beyond social communication and interpersonal relationships, socio-affective competence has a large impact on human collaborative and organizational behaviors and their outcomes. For example, the ability to establish and reciprocate trust is fundamental to successful and effective collaboration [173]. Therefore, HRI researchers working on human-robot collaboration, such as applications in workplaces, can also benefit from improving the socio-affective competence of robots. For instance, a previous study has found that a robot able to express emotions and use appropriate gaze behaviors can inspire cooperative and adaptive behaviors in its human teammates, which resulted in more successful and efficient collaboration compared to robots without emotional expression or with random gaze behaviors [69]. Thus, in this section we review socio-affective competence in the context of human prosocial behaviors to support better design of human-robot collaboration systems.

The impact of socio-affective competence on collaborative behaviors has been studied intensively in behavioral economics and economic games, mainly in the context of social norms [45]. Social norms are external, socially defined normative standards which are the most socially appropriate action for a decision maker in a given set of circumstances [44]. Adherence to social norms is tied to a person's social reputation and perceived prosociality [4]. Social norms vary across cultures and individual sensitivity to social reputation. They are constantly changing, following the shifting expectations of the majority in a social group [87]. Adherence to, or violation of, social norms has mainly been studied in the context of three economic games: the public goods game [96], the trust game [16], and the dictator game [80]. The public goods game measures people's prosociality, and investigates how awareness of social reputation motivates prosocial behaviors [45]. The trust game measures trust and reciprocity [24], which accommodated indicators of cooperativeness and adherence to social norms [164]. The dictator game measures people's adherence to the fairness social norm [40].

Collaboration and prosociality	
Socio-affective competence attributes	**Assessment methods**
Ability to enlist others in collaborative task	Subjective, e.g., [144]
Awareness of privacy of self and others	Subjective, e.g., [30]
Sensitivity to social reputation and perceived prosociality	Objective, e.g., [75]
Ability to establish and maintain social reputations	Subjective, e.g., [144]
Ability to comply and cooperate to maintain reciprocity and fairness in collaboration	Objective, e.g., [40]
Ability to establish and reciprocate trust	Objective, e.g., [24]

Table 3. Socio-affective competence attributes related to collaboration and prosocial behaviors.

Variations of these games have been applied to evaluate the perceived prosociality of robots and a user's trust towards a robot (e.g., Wu et al. [180]).

An interesting aspect exposed by behavioral economics studies is the temporal dimension of social errors. Trust requires time to cultivate. However, established trust can be violated by short instances, and the difficulty and time required for trust recovery depends on various factors, such as compensation and repair strategies [38]. This indicates that the perceived socio-affective competence of a robot and social errors in HRI should also be investigated in longitudinal settings beyond short-term, single interaction sessions.

Our review of human collaborative and organizational behaviors allows us to zoom out from interpersonal social interactions to collaborations within and across groups. This leads to additional attributes of socio-affective competence which we report in Table 3.

We expect that researchers working on human-robot collaboration may follow the identified attributes and improve the perceived socio-affective competence of a robot, which potentially improves the team dynamics and outcome of collaborative tasks. For example, in peer tutoring, when the tutees have privacy or trust concerns toward their tutor, they can feel demotivated to participate in learning activities [5]. Thus, an educational robot perceived as trustworthy and privacy-preserving may inspire stronger engagement and motivation in its students, resulting in better learning outcomes.

2.5 Attributes of socio-affective competence in human interpersonal communications

In previous sections, we first reviewed what functions emotions serve in human cognitive and communicative processes to provide theoretical background for designing HRI systems capable of emotional interactions. We then reviewed the convergence of emotions and behaviors in interpersonal communication, i.e., empathy and entrainment, which often predict the social dynamics and outcomes of a social interaction. The review of emotions and empathy results in a set of socio-affective competence attributes related to emotional interaction and social skills, which we reported in Table 1. A person's capability to engage in reciprocal social interaction and communication may be impaired, and various evaluation approaches have been used in the mental health context to measure such impairment. We reviewed these assessments and extended the set of socio-affective competence attributes to include dimensions related to understanding others and communicative functions, which we reported in Table 2. Finally, we zoomed out to the societal and organizational level, and reviewed social norms, prosociality, and human collaboration to identify relevant socio-affective competence attributes, which we reported in Table 3.

Combining Tables 1, 2, and 3, we identify thirty major attributes of socio-affective competence under five categories, which people use to evaluate social interaction and their social relationship with others. We expect these attributes to be applicable to a broad class of HRI scenarios and application domains. Thus, they form the basis of our taxonomy of HRI social errors, which will be presented in Table 4 of Section 4. These attributes can be used to decompose how an HRI social error influences perceived socio-affective competence and social relationship, and in turn guide the design of HRI systems that are capable of natural and socially acceptable interaction. With better understanding of people's perceptions and needs regarding social interaction, HRI researchers can adopt a more user-centered design perspective. This yield a potential boost to interaction outcomes. For example, individuals participating in robot-assisted gait training may respond better to instructions from a robot that exhibits empathic concerns, resulting in better health outcomes. Moreover, these attributes can be applied to evaluate the perceived socio-affective competence of such systems and their social relationship with users. Our review exposed that in human studies, a number of socio-affective competence dimensions rely on subjective measurement, such as self-reported questionnaires used to assess assertiveness. To evaluate a robot's socio-affective competence on these dimensions, novel objective measurements need to be developed in future HRI research.

Note that for each attribute of the socio-affective competence we identified, the judgment of whether a behavior is normative or not varies across cultural, contextual, and individual factors. Take the ability to express emotions as an example. Different cultures have different norms regarding what is an "appropriate" emotional expression. Our list of attributes summarizes the *dimensions* along which people judge the socio-affective competence of themselves and others. However, the perception of socio-affective competence is flexible and context-dependent. Thus, the discussion of whether a specific behavior is deemed as having positive or negative impact on the perceived socio-affective competence and the degree of such impact should be situated in specific scenarios.

3 RELATED WORK ON HRI ERRORS

In Section 2, we focused on existing human studies on social interaction, and identified thirty main attributes of socio-affective competence under five categories. Because of the complexity in human social interaction, a robot designed to interact with people is bond to make errors. The mean time between technical failures for robots in the wild is often less than a few hours, and current HRI systems are not yet mature enough to effectively handle unexpected events [70]. Therefore, HRI errors have attracted growing interest in the research community, with recent studies focusing on data collection and experiments contributing to the understanding of how errors influence a user's perceptions and behavioral responses, as well as effective recovery strategies [90]. We survey existing studies on HRI errors with a user-centered perspective in Section 3.1. We then review existing taxonomies of HRI errors in Section 3.2. This allows us to plot an overview of the current research landscape, and where our taxonomy of HRI social errors stands.

3.1 Impact of HRI errors

3.1.1 HRI errors and a user's trust.

Previous analyses of HRI errors' impact on a user's perception have mainly focused on their influence on users' trust towards the HRI system. For example, Tolmeijer et al. [171] proposed a taxonomy of trust in HRI and strategies to mitigate breaches. Trust has direct impacts on human's decision-making, and willingness to accept information or follow instructions from robots [64]. Three main aspects of trust have been analyzed in current HRI studies: the environment perspective, the human perspective, and the robot perspective [157].

The environment perspective of trust studies characteristics of the task and their influence on trusting behaviors in HRI. For example, risk level [145] or interdependency between team members [107]. The human perspective studies demographic characteristics of users and their trusting behaviors, such as personality [112] or ethnicity [140]. User's mental and cognitive states have also been analyzed, such as workload [17]. The robot perspective studies features of a robot and their influence on a user's trust. In terms of appearance of a robot, it has been identified that people tend to trust embodied agents more than non-embodied agents [174]. The attractiveness of physical appearance and voice of agents was also shown to influence trust in user-agent interaction [184]. In terms of capability of a robot, erratic behaviors of robots were shown to negatively influence people's trust towards them [177]. Additional to trust, other attributes of a user's perception can be influenced by erratic behaviors of a robot, such as perceived naturalness of the HRI experience [73]. Note that erratic behaviors of a robot do not necessarily lead to observable changes in interaction behaviors or compliances in short-term HRI [156]. This may be related to interaction contexts or individual differences, i.e., the environment perspective and the human perspective.

Beyond changes in a user's trust caused by the environment perspective, the human perspective, or the robot perspective of HRI, how a robot reciprocates a user's trust may also influence the interaction. In both social and non-social situations, humans can feel betrayed when their trust is not reciprocated [81], or when their expectation is unsatisfied [55]. Such a feeling leads to decreased willingness to comply and increased willingness to punish when the desired outcome is not achieved [2]. Thus, a robot's ability to inspire trust in its user may also influence the effectiveness of its error recovery strategies.

3.1.2 The performance, social, and temporal aspects of HRI errors.

The performance aspect and social aspect of HRI are often not distinguished in current HRI studies. For example, Rossi et al. [148] listed a set of verbally described HRI scenarios in which a domestic service robot exhibits erratic behaviors, and collected people's perception of the severity of these errors through crowd-sourcing. These scenarios are a mix of performance errors and social errors by our definition, and Rossi et al. [148] have reported both types of error being perceived as severe.

To the best of our knowledge, the only recent HRI error study that distinguished social errors and performance errors is the work of Giuliani et al. [56]. The authors analyzed user behaviors in multiple HRI studies when four types of failures in the communicative function of robots occurred, namely long dialogue pauses, repetitions in the dialogue, misunderstanding, and complete discontinuation of the interaction. They classified these failures into two classes: technical failures, and violations of social norms. In particular, they studied two types of violations of social norms: a robot executing interaction steps at the wrong time, or a robot showing unusual social signals during dialogues. Giuliani et al. [56] focused on the dialogue aspect of social interactions and defined social signals as verbal and non-verbal signals humans use in a conversation to communicate their intentions. To understand users' verbal and non-verbal reactions to HRI errors, they collected manual annotations of user behaviors on five modalities, namely speech, head movements, hand gestures, facial expressions, and body movements. Giuliani et al. [56] compared the mean number of occurrences of each type of user behavior per interaction in social norm violation versus technical failure. Their analyses showed different behavioral patterns of the participants when different types of error occurred. In particular, participants talked more when a robot exhibited social norm violations compared to when it exhibited technical failures. In their further investigations [115, 116], they also found that participants showed more social signals when robots violated social norms compared to when they exhibited technical failures.

The work of Giuliani et al. [56] indicated that distinguishing social errors and performance errors provides additional insights to how users react to HRI errors. However, their studies have two major limitations: On the one hand, their analyses have focused on the communicative aspect of dialogue-based social interactions. Our review in Section 2 has shown that social interaction is more than verbal and non-verbal communication. It contains other aspects such as emotional states and social relationship. Thus, their approach of categorizing social errors as violations to two types of social norms in dialogue is limited. On the other hand, their work focused on audiovisual analyses of users' behaviors. How HRI errors influence users' perceptions, such as trust, is yet to be understood. Our taxonomy aims at providing a broader view of social errors in HRI and their influences.

Beyond analyzing how HRI errors influence a user's perceptions and behaviors, another type of HRI error studies focused on recovery strategies and managing users' expectations. For example, Sebo et al. [159] investigated a robot's trust violation and repair during a competitive game with a participant. They found that when a robot broke a promise previously made to a participant, the participant indicated lower trust towards the robot and exhibited retaliation in later stages of the game. They studied two types of repair strategies addressing the trust violation, namely apology and denial, and their results suggested that apology was more effective. In another example, Lee et al. [97] compared different recovery strategies on reducing the negative consequence of breakdowns in robotic services. They found that users with a relational orientation during HRI responded best to apologies, while those with a utilitarian orientation responded best to compensation. This indicates the importance of understanding the users' interaction goals when designing and evaluating an HRI system. Lee et al. [97] made the distinction between social and performance oriented interaction goals of a user, and discussed how the effect of a robot's social behaviors differ for these two types of interaction goals. The service breakdown (bringing a wrong drink) in their simulated study is a performance error by our definition, while a robot displaying inappropriate social behaviors (e.g., using an ineffective recovery strategy) after a service breakdown is a social error by our definition. Thus, such analyses of users' expectation and reception to different recovery strategies can be considered as HRI social error studies, which aligns with the "Self-consciousness and anticipation of others' reception of self" attribute of socio-affective competence in Table 2.

Beyond the performance and social aspects of HRI errors, the temporal aspect can influence a user's reception and perception of a robot as well. Most existing HRI error studies are short-term, single session HRI experiments. Long-term, in-the-wild HRI studies are extremely scarce [74]. In a representative study on understanding users' abandonment of robots [32], 70 rabbit-shaped robots were placed in participants' homes for six months. De Graaf et al. [32] analyzed participants' acceptance of the robots using questionnaires and interviews at 6 data collection sessions ranging from 2 weeks before to six months after the introduction of the robots. They found that the robot's lack of adaptability and sociability were the two determinants of users' abandonment. This study indicates the importance of a robot's perceived socio-affective competence in long-term HRI. Similarly, Beane [14] analyzed the usage of telepresence robots deployed at a commercial hospital. It was found that the robots were abandoned and stowed away shortly after deployment, as they only served as a symbol of technological novelty and offered limited interaction and use cases. To enable personalization and adaption during HRI, it is important to involve end users at the system design stage. For example, Kubota et al. [92] included target users in participatory

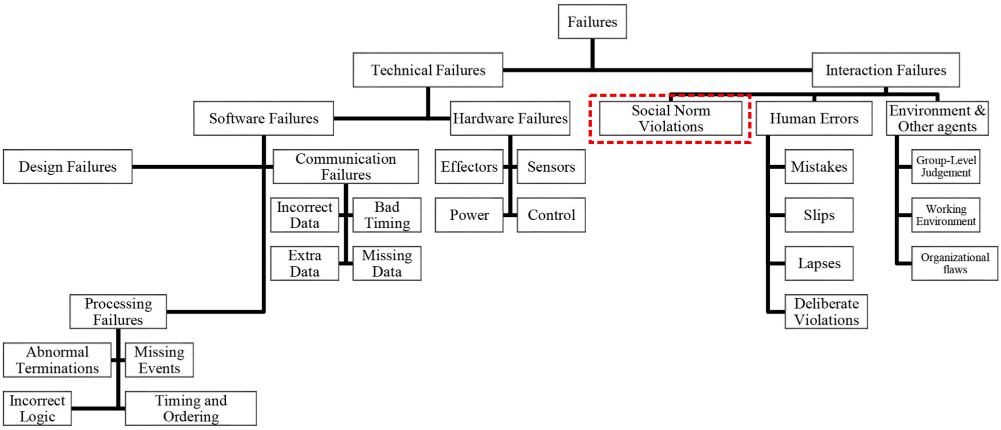

Fig. 1. The human-robot failure taxonomy proposed by Honig and Oron-Gilad [70].

design workshops for programming functions of a social robot for neurorehabilitation. Their results demonstrated benefits in both usability and health outcomes of the robot.

3.2 Existing taxonomies of HRI errors

To better facilitate systematic analysis of errors in HRI, several taxonomies of HRI errors have been proposed. In a recent effort, Honig and Oron-Gilad [70] proposed to classify HRI errors as technical failures and interaction failures. However, as shown in Figure 1, compared to other sub-categories of HRI errors, social norm violation requires further analyses and more detailed classification. In addition to Honig and Oron-Gilad [70], HRI error taxonomies have focused on the functional aspect of HRI, rather than users' perceptions of these errors. Here we list representatives of such function-focused HRI error taxonomies:

- Laprie [94] classified errors by severity as benign errors and catastrophic errors;
- Ross et al. [147] classified errors by recoverability as anticipated errors, exceptional errors, unrecoverable errors, and socially recoverable errors (the robot can continue on with the original plan with appropriate assistance from its environment);
- Carlson and Murphy [23] classified errors into physical errors and human errors first, then further classified physical errors by severity and recoverability, and classified human errors as design errors and interaction errors;
- Steinbauer [168] classified errors as interaction errors, algorithm errors, software errors, and hardware errors;
- Brooks [19] classified errors as communication errors and processing errors.

These existing taxonomies of errors in HRI classified errors either by their causes and responsible functional components, or by the magnitude of their consequences and the difficulty in recovering from them. The influences errors may have on a user's perceptions of the system or their interaction behaviors have been overlooked. Compared to previous work, our proposed HRI error taxonomy provides a user-centered perspective instead. Moreover, these existing taxonomies have focused on the functional dimension of HRI, while the social-emotional dimension of HRI requires further analyses. On the contrary, our taxonomy of social errors in HRI focuses on how errors in HRI influence users' perception of the socio-affective competence of a robot and their perceived social relationship with it. This is especially important for HRI systems with socially oriented goals, i.e., social robots.

4 TAXONOMY OF SOCIAL ERRORS IN HRI

As discussed in Section 3, a systematic and user-centered understanding of errors on the social-emotional dimension of HRI is required. This motivates us to propose a taxonomy of social errors in HRI. In particular, inspired by the thirty main attributes under five categories that humans adopt to perceive socio-affective competence and social relationships (see Section 2), we classify social errors in HRI into five categories, and thirty sub-categories in Table 4. Each category contains both instantaneous social errors which may occur in short-term HRI, and social errors specific to long-term HRI. Definition and evaluation approach for each of these attributes in human studies were listed in Tables 1, 2, and 3.

This proposed taxonomy illustrates the breadth and complexity of social errors in HRI. Note that the error types are not entirely independent to one another and can be correlated. For example, a robot's failure in maintaining engagement of its user during interactions (E4.5) can be a result of it generating irrelevant or incomprehensible replies to verbal or non-verbal communications of the user (E4.2).

Because human perception of a breach and its severity is flexible and context-dependent. To interpret the proposed taxonomy more thoroughly and to illustrate different types of social errors, we situate our discussion in HRI scenarios. Scenario-based design is a widely used approach in human-computer interaction studies [150]. The core idea is to use narrative description of envisioned usage episodes to guide a concrete yet flexible discussion. Scenario-based design emphasizes people and their experiences, which aligns with the user-centered perspective of our taxonomy of HRI social errors. We create three hypothetical HRI scenarios following the ecological validity guideline, i.e., these are believable situations which approximate real-life situations. They are representatives of a number of existing HRI studies under similar contexts, as detailed below:

- *Olivaw* is a robot receptionist at the front desk of the Faculty of Information Technology, Monash University (Figure 2a). A human receptionist, Sid, is also visible and accessible to visitors. Visitors to the front desk may either be first timers, or staffs and students of the department who visit the reception desk regularly. Olivaw carries out receptionist duties, such as answering questions or handing out marked assignments. Other than common knowledge about the department, it also has access to private information such as exam marks. Example HRI studies similar to this robot receptionist scenario include Nuccio et al. [125] and Guo et al. [61].
- *Calvin* is a robot carer for Frank, a 75 year old granddad living alone in his countryside house. Frank has dementia and his condition is progressing over time, but he refused his kids' suggestion of moving to a care home. Calvin helps Frank with daily activities, such as making breakfast or reminding him to take his medicines. It also encourages Frank to carry out activities benefiting his mental and physical health, such as picking up gardening as a hobby and joining a local book-reading club.[2] Example HRI studies similar to this robot carer scenario include De Graaf et al. [32] and Sabelli et al. [155].
- *Reventlov* is a robot research assistant at the Davis station, one of the three Australian permanent bases and research outposts in Antarctica (Figure 2b). The Davis station is operational all year round, and typically has 120 personnel in summer time, and 18 in winter. The Davis station contains various living facilities for the expeditioners, such as a communal dining room and a gym. Personnel at the Davis station carry out research projects, for instance analyzing the impact of environmental changes on the Antarctic marine ecosystems. They also take turns to contribute to the day-to-day running of

[2] Some characters were inspired by the 2012 movie *Robot & Frank*: https://www.imdb.com/title/tt1990314/

E1. Breach in empathic and emotional reactions (Section 4.1)	
Short-term	E1.1. Incorrect or absent recognition of a user's emotions
	E1.2. Showing inappropriate or absent emotional expressions during interactions
	E1.3. Exhibiting unstable or unusual changes in emotional behaviors
	E1.4. Unresponsive or misaligned responses to a user's emotions
	E1.5. Inappropriate or absent entrainment responses during interactions
Long-term	E1.6. Failures in perspective-taking to establish mutual understanding
E2. Insufficient social skills (Section 4.2)	
Short-term	E2.1. Incorrect or absent recognition of the social context or social relationship
	E2.2. Failures in adjusting behaviors according to social contexts, or exhibiting inappropriate behaviors misaligned with the social relationship with a user
	E2.3. Exhibiting inappropriate level of independence or assertiveness
	E2.4. Violation or over-adherence to social norms
	E2.5. Failures in initiating or responding to reciprocal social communication
	E2.6. Failures in adhering to social routine
Long-term	E2.7. Failures in establishing or maintaining social relationships
E3. Misunderstanding the user (Section 4.3)	
Short-term	E3.1. Incorrect or absent assessment of a user's knowledge-states
	E3.2. Incorrect or absent assessment of a user's intentions
	E3.3. Unresponsive or misaligned responses to joint attention and eye gaze
	E3.4. Unresponsive or misaligned responses to shared enjoyment
Long-term	E3.5. Incorrect or absent assessment of a user's anticipation or reception of HRI
E4. Insufficient communicative functions (Section 4.4)	
Short-term	E4.1. Failures in initiating verbal or non-verbal communications
	E4.2. Synthesizing irrelevant or incomprehensible replies to verbal or non-verbal communications
	E4.3. Synthesizing inappropriate or absent non-verbal expressions
	E4.4. Misalignment in turn-taking
Long-term	E4.5. Failures in maintaining engagement of a user during interactions
	E4.6. Absence of adaption or exhibiting incoherence during interactions
E5. Breach in collaboration and prosociality (Section 4.5)	
Short-term	E5.1. Failures in enlisting collaborators
	E5.2. Violation of privacy of a user or others
	E5.3. Being insensitive to or damaging social reputation of a user or others
Long-term	E5.4. Failures in establishing or maintaining social reputations of self
	E5.5. Failures in maintaining reciprocity or fairness in collaboration
	E5.6. Failures in inspiring or reciprocating a user's trust

Table 4. Taxonomy of social errors in HRI.

the station, e.g., cooking. Reventlov collaborates with researchers in scientific tasks, for instance retrieving and documenting samples. It also assists with station maintenance, such as shoveling snow on the premises. Example HRI studies similar to this robot research assistant scenario include Seo et al. [160] and Vannucci et al. [175].

We chose these three HRI scenarios to investigate different aspects of interactions: In Olivaw's scenario, the robot interacts with various people with different backgrounds and goals. Their interaction with Olivaw varies in length and frequency and they have different

(a) Front desk of the faculty.

(b) The Davis Station [37].

Fig. 2. Illustration of the hypothetical HRI scenarios.

social relationships with Olivaw. The goal of Olivaw is to provide services, such as question-answering, as well as to facilitate engaging social interactions to make the interactees feel welcomed and at ease; In Calvin's scenario, the robot interacts with one specific user in a longitudinal setting. The interaction also needs to be designed with additional attention to address the mental health context of the user. The goal of Calvin is to support Frank's physical and mental well-being through long-term interaction; In Reventlov's scenario, the robot participates in collaborative tasks of various risk and complexity levels. The goal of Reventlov is to achieve task success in individual collaborations, as well as to form a productive human-robot team with positive team dynamics over time.

Recall that our proposed taxonomy of HRI social errors is derived from dimensions of socio-affective competence drawn from different areas: E1 (errors in emotional interaction) and E2 (insufficient social skills) were motivated by assessments of social skill and social relationship of the general population, E3 (misunderstanding the user) and E4 (communicative errors) were motivated by diagnostic evaluation of reciprocal social interaction in the mental health context, E5 (errors in collaboration) was motivated by collaboration and prosocial behaviors. Thus, in our three hypothetical HRI scenarios, E1 and E2 may be more relevant to Olivaw's scenario, E3 and E4 may be more relevant to Calvin's scenario, and E5 may be more relevant to Reventlov's scenario. However, these socio-affective dimensions are connected to each other and together they cover the multiple facets of social-emotional interaction, as discussed in Section 2.5. Thus, these social errors in HRI are applicable to application scenarios beyond the specific areas that they originated from, and all five types of social errors may occur in an HRI scenario. In the following sections, we will discuss the proposed taxonomy of HRI social errors situated in these three HRI scenarios, as well as relating to existing HRI studies.

4.1 Breach in empathic and emotional reactions

Empathic and emotional reactions may not always be necessary in HRI [109]. Regarding Olivaw's scenario, in short-term interaction sessions, without emotional reaction functions, it can still provide basic services such as giving directions to visitors. However, without proper monitoring of a user's emotional states (E1.1 emotion recognition errors), it will be difficult to design Olivaw to provide personalized and engaging interaction experiences. For example, if Olivaw follows the exact same script for introducing the department without adjusting to a visitor's interests (E1.4 unresponsive to user's emotions), it may soon lose its audience's attention. In fact, a recent study found that a dialogue system unresponsive to a user's inferences was rated as significantly worse on user engagement and user experience than an appropriately tailored system [183]. Other than task-oriented interactions, a receptionist's interaction with visitors often serve social functions. When people interact with a receptionist robot, they expect social interactions in addition to question-answering [125]. Thus, Olivaw needs to address these expectations in order to provide satisfactory services. In particular, failures in displaying emotional synchronicity and entrainment with its dialogue partner (E1.5) may result in lower rapport between a user and Olivaw [68].

Perspective taking and mutual understanding (E1.6) is closely related to a recently proposed HRI notion "Affective Grounding" [78]. Affective grounding refers to coordination of affects in interaction with the purpose of building shared understanding on how behaviors should be interpreted emotionally and what counts as "appropriate". Failures in establishing mutual understanding and affective grounding (E1.6) can result in misjudgment of the goal of a task or inaccurate estimation of the appropriateness of an action or response, especially in long-term HRI.

In Calvin's scenario, a breach in empathic and emotional reactions is particularly harmful because of the vulnerability of its user. In individual short-term interaction sessions, when Calvin is not able to accurately recognize Frank's emotional states (E1.1) or exhibit empathy and entrainment (E1.5), providing help and mental health care in a timely fashion becomes nearly impossible. As a long-term companion, Calvin is expected to acquire mutual understanding and common grounding with Frank over longitudinal interaction [154]. Failures to do so (E1.6 unable to establish mutual understanding) may cause difficulty in building a social relationship between Frank and Calvin over time.

Through longitudinal interaction, a robot can learn to improve its ability to recognize emotions of a specific person, and that person can better recognize the synthesized expressions of the robot, i.e., mutual adaption. Such learned adaption enables a robot to respond with empathy [131], and over time establish a companionship with its user [58]. Because the goal of Calvin is to provide longitudinal service to Frank, its socio-affective functions need to be personalized through accumulated interaction experiences. Furthermore, adaption to the progress of Frank's dementia is required for Calvin's support to remain effective over time. It is also important to take into account that Frank can exhibit negative or labile emotions because of his dementia [139]. Thus, instead of always converging towards Frank's emotional states, a more productive strategy for Calvin may be to maintain a positive and stable emotional state for Frank to converge with. On the contrary, if Calvin appears unstable in its emotional states (E1.3), it may worsen Frank's condition. Current studies on using robots for dementia care are extremely limited in incorporating personalization and adaption [119]. Thus, more longitudinal HRI research is required for a better understanding.

In Reventlov's scenario, as discussed in Section 2.1, emotion is an important communicative cue which humans use to express their mind and interpret other's. Thus, when Reventlov is unable to recognize emotions of its human teammates in an accurate and timely manner (E1.1), its ability to assess joint intention of the team can be impaired, while such ability is key to effective and successful human-robot collaboration [13]. Similarly, failures in Reventlov's emotion synthesis functions (E1.2) can result in difficulties for its collaborators to perceive its intentions. In addition, mutual understanding and common grounding is critical for human collaborative problem solving [41]. Thus, failures in establishing mutual understanding (E1.6) create obstacles for team integration and lower the efficiency of the collaboration in the long term. Misjudgments in human-robot collaboration may also result in negative impact on the emotional states of humans in the team [13]. Note that the polar expedition context puts the human-robot team in an extreme environment. Humans can experience psychological changes caused by such a long period of isolation and confinement, including impaired cognitive ability, negative affect, and interpersonal tension and conflict [133]. Thus, the negative impact of Reventlov's errors in empathic and emotional reaction may be magnified over time under such circumstances.

As discussed here, in short-term HRI, a breach in empathic and emotional reactions can lead to user engagement breakdowns or failures in achieving the goal of the interaction session. However, violation of existing expectancy may also result in humor [36], and a breach in the interaction norms may instead amuse a user and increase the memorability and enjoyment of an HRI session. Although such effect may disappear when errors occur repeatedly and when the novelty effect [84] has worn off. The induction of humor can be difficult to control and the outcome can vary for different individuals. Thus, impact of the same social error may differ in short-term vs. long-term or repeated HRI, and researchers should take the temporal dimension of HRI into account when studying social errors.

4.2 Insufficient social skills

Social skills facilitate human interpersonal relationship building and managing, as discussed in Section 2.2. Thus, insufficient social skills are likely to damage the perceived social relationship between a user and a robot. In Olivaw's scenario, it provides services to a heterogeneous population. Thus, not incorporating interaction history and frequency of visits may reduce a user's satisfaction towards its service (E2.1 absent recognition of social context). For example, a person visiting the reception desk for the first time will be expecting a lower intimacy level in their social interaction with Olivaw compared to a member of the department who is a regular visitor. If Olivaw displays behaviors mismatching the perceived intimacy level (E2.2 errors in adjusting behaviors according to social contexts), for instance standing too close to a first-time visitor while having a conversation, it may induce feelings of unease in the person and arouse more rejections or conflicts during the interaction [6].

Besides social context, heterogeneity in the demographic characteristics of the visitors also requires Olivaw to adjust its knowledge on social norms and social routines to reciprocate social communication appropriately. For example, a Chinese visitor may ask "Have you eaten yet?" as a common greeting equivalent to "How are you?", and her culture's social routine does not expect a detailed answer to this question. If Olivaw gives a lengthy response about its battery charging process, it may confuse the person (E2.6 failures in adhering to social routine). Regarding expression of assertiveness, a close relationship between assertiveness and confidence was found in human interactions [86]. Thus, failures in displaying assertiveness and independence (E2.3) may reduce users' perception of Olivaw's confidence in providing required services.

In Calvin's scenario, Heerink et al. [67] analyzed short-term interactions between elderly users and robots, and found that the participants felt more comfortable and were more expressive during interaction when the robot followed normative social behaviors and engaged in reciprocal social communication. Thus, if Calvin fails to initiate or respond to reciprocal social communication (E2.5), it may lead to Frank's rejection of Calvin or compromise his desire to use it. Beyond short-term interaction, it is important for Calvin to establish and maintain a social relationship with Frank during longitudinal interaction. Social relationship has known health benefits [72], and perceived social relationship with another entity represents the perceived adequacy of resources provided by the other entity [26]. Thus, if Calvin is unable to establish and maintain a social relationship with Frank over time (E2.7), it is unlikely to achieve the expected health benefits of engaging Frank in social interactions, and Frank may not rely on Calvin for physical or mental support. For example, after a period of time, Frank may become less willing to chat with Calvin when it prompts a conversation, or follow Calvin's instructions in doing a physical exercise.

Unlike Olivaw, as a personal robot, Calvin is expected to start with a general knowledge on social norms and social routines based on demographic characteristics of Frank, and then gradually tailor this knowledge to suit Frank's preferences. However, this does not equal to designing a robot which is always agreeing and providing shallow flattering. Considering Frank's dementia, he may exhibit violation or over-adherence to social norms during his interaction with other people, which can harm his social relationship with them. For example, as discussed in Section 2.3.2, dementia patients can be insensitive to other's embarrassment or privacy [30]. Thus, it is important for Calvin to maintain a suitable level of independence and assertiveness (E2.3 inappropriate level of independence) to support Frank in interacting with others.

In Reventlov's scenario, failures in establishing a social relationship with its human collaborators (E2.7) can degrade their trust towards it [77]. A lack of trust can be especially harmful in the polar expedition scenario due to the high-risk nature of some of the tasks and the extreme physical environment [145]. Beyond managing its social relationship with collaborators, Reventlov is also expected to be aware of the social relationships between the collaborators to achieve better team outcomes. For example, in a previous study on human-robot collaborative problem-solving, a robot's intervention on conflicts between human team members has been shown to increase the team's awareness of the conflict, motivating the team to regulate its behaviors to process the conflict [79].

Personnel at the polar station often consists of people from various backgrounds with different expectations on social norms and social routines. During teamwork, reciprocal social communication is an important lubricant for a productive and effective collaboration [53]. Thus, failures in adjusting actions based on specific collaborators (E2.2) may limit Reventlov's ability to enlist and collaborate with its teammates. In addition, assertiveness is essential to effectively communicate action plans to the team [63]. Failures to exhibit assertiveness (E2.3) can reduce other team members' evaluation of the capability of Reventlov and their willingness to comply with plans it proposes. However, this does not equal to being assertive all the time. For example, when the robot makes a mistake, it is expected to adhere to the social norm of admitting its fault and apologizing to its teammates (E2.4 violation of social norms).

4.3 Misunderstanding the user

Understanding users' intentions allows a robot to operate proactively [142]. Failures to do so can lead to less efficient services or unsatisfactory task outcomes. In HRI research, human intention inference has been studied to enable autonomous robots to plan their paths around people safely [83], or to engage in mutual action planning, such as handing over objects [123]. User understanding is also the foundation of personalization in HRI. A recent review of user profiling and behavioral adaption in HRI has identified that a desirable HRI should be adaptive on three aspects: physical, cognitive, and social [149]. These aspects are interdependent: physical adaption addresses user-specific behaviors, cognitive adaption addresses users' intention of such behaviors, and social adaption addresses relationship between a user and a robot which changes over physical and cognitive adaption. Personalization can happen both in a passive manner where a robot learns by observation (e.g., Kato et al. [82]), or in a proactive manner where a robot engages the user in teaching it information required for personalization, i.e., a teachable robot (e.g., Churamani et al. [25]). Personalized and adaptive HRI is still in its infancy, especially for longitudinal HRI [74].

In Olivaw's scenario, a lack of understanding of its users' intentions and knowledge-states can limit its ability to provide services. For example, when Olivaw is giving directions to visitors attending a seminar, members of the department may only require the room number, while visitors from other departments may require more detailed instructions (E3.1 absent assessment of a user's knowledge-states).

Regarding assistive technology, intention inference has been studied to ensure a safe and adaptive support [141]. When there is possible impairment in users' cognitive abilities, a robot's ability to provide proactive support becomes especially important. In Calvin's scenario of long-term HRI, compared to short-term HRI, there is potential for learning a deeper level of understanding of its user through longitudinal interaction. Such understanding is crucial for establishing social engagement and developing social bonds [22]. If Calvin cannot understand and adjust to Frank's reception of different stages of treatment plans, it will not be able to provide effective care over time (E3.5 incorrect assessment of a user's reception of HRI). Failures in assessing the intention and attention of a user can cause a robot to ignore service requests too, which can result in safety concerns in Calvin's scenario. For example, imagine Frank is having a stroke and lost his ability to speak or walk, so he points to an emergency button on the wall next to the bookshelf for contacting his carer, expecting Calvin to press it for him. However, Calvin mis-interprets his intention and brings a book back instead (E3.2 incorrect assessment of a user's intention). This may in the end cause delay in Frank's treatment.

In human communication, eye gaze serves important functional and social roles. People were found to be sensitive to gaze behaviors of robots and they felt more engaged with a robot that establishes mutual gaze [89]. Thus, unresponsiveness to eye gaze or joint attention (E3.3) can result in a robot being perceived as distracted or less socially friendly. Considering Frank's dementia, his eye gaze behaviors may differ from normative behaviors. However, most existing models for human behavioral analytics are based on normative behavior data collected from a healthy population. Thus, Calvin needs to adjust its recognition models to better suit Frank's needs. In terms of shared enjoyment, it represents a pleasure in interactive participation and *fun*, which is important for relationship development [47]. Thus, the design of a companion robot requires attention to shared enjoyment. For example, if Calvin is helping Frank with gardening and does not respond to his excitement during this co-creation experience (E3.4 unresponsive to shared enjoyment), Frank may have a worse mood and become less motivated to participate in the future.

Regarding human-robot collaboration, a robot's ability to assess the knowledge and intention of its team members has been found to benefit the objective and perceived performance of the collaboration [104]. Failures to do so can damage the collaboration efficacy on action planning and resource allocation (E3.1 incorrect assessment of a user's knowledge-state or E3.2 intentions). As for responding to shared enjoyment and building rapport, it has been identified that in professional human collaborations, a good rapport is beneficial in terms of functioning of the team and overall team satisfaction. People expect a social relationship with their teammates, whether they are humans or robots [160]. Thus, in Reventlov's scenario, it is desirable to address such expectations and design Reventlov to establish rapport with its teammates.

4.4 Insufficient communicative functions

Verbal and non-verbal communication has been the center of decades of studies on human social interactions [98]. Thus, communicative function of a robot is critical to its ability to engage in social interaction with humans. Note that communicative functions of a robot are not limited to language abilities, such as speech recognition and synthesis. For example, Paro the robotic seal has no verbal speech synthesis function and only responds to limited spoken keywords, yet it is able to effectively engage its users and achieve positive health outcomes in elderly care and mental health care [154]. Previous studies have identified that human social communication is essentially multimodal [118]. Therefore, it is desirable for HRI systems to have the capability of engaging in multimodal communication with users [49].

Regarding mutual adaption in communication, Oviatt et al. [129] analyzed automatic speech recognition for conversational interface. They found that users exhibit hyperarticulation when repeating speech for error resolution, such as changing the speed and amplitude in their speech. This can cause a downward spiral in error scenarios, i.e., the hyperarticulated speech becoming even harder for the speech recognizer to register. Similar mutual adaption may exist in HRI when a robot exhibits insufficient communicative functions. However, there has been limited research on mutual adaption to HRI social errors.

Note that for a conversational agent, holding a natural and engaging dialogue is one of its main objectives. Thus, insufficient communicative functions may degrade a user's perception of a conversational robot's competence in achieving tasks as well as degrading her perception of the robot's socio-affective competence. In this case, insufficient communicative functions is a social error and a performance error at the same time. In Olivaw's scenario, because the majority of its tasks are conversational, confusions caused by insufficient communicative functions can significantly decrease its ability to provide satisfactory services, as suggested by previous research [56]. For example, if Olivaw fails to answer a visitor's question in a comprehensible manner (E4.2 irrelevant or incomprehensible replies), or if it has long delays during a dialogue (E4.4 turn-taking errors), a visitor may become disengaged (E4.5 failure to maintain user engagement) and aborts the interaction.

In Calvin's scenario, considering Frank's dementia, it is important for Calvin to communicate in a reliable, precise, and easy-to-interpret manner [42]. For example, using over-articulation in the synthesized speech [166], or including more pronounced gestures and expressions during the communication [182]. Individuals with dementia can feel socially isolated and experience a depressed mood because of the progressive impairment in their cognitive functioning [139]. Thus, a key component of dementia care is maintaining communication [178]. Patients with mild or moderate dementia are often able to engage in social communication more verbally than those with severe dementia [58]. Therefore, multimodal communication such as touch, facial expressions, and non-speech audio become increasingly important as the disease progresses. On the contrary, if Calvin fails to tailor its expressions to Frank's condition (E4.3 absent non-verbal expressions), it may become difficult to engage Frank in social communication. This can result in negative influence on Frank's dementia, although long-term HRI studies are required to understand the magnitude of such influence. Furthermore, because Calvin's interaction with Frank is longitudinal, if it does not learn from the interaction history and maintain coherence (E4.6 absence of adaption), Frank may stop interacting with it after a short period [32].

In Reventlov's scenario, verbal and non-verbal communications allow the team to establish common ground and adjust actions in a timely fashion [18]. Thus, insufficient communicative functions of Reventlov, such as failing to explain a plan clearly before a mission begins (E4.1 errors in initiating communication), may reduce the efficiency and robustness of the human-robot team.

4.5 Breach in collaboration and prosociality

As discussed in Section 3.1, trust has been a focus of current HRI error research. Note that performance or social errors in a short-term interaction session may not necessarily lead to degraded trust of a user. Additionally, there has been evidence arguing that degraded trust may not influence people's decision to follow a robot's instructions in short-term HRI [54]. However, in long-term interactions, when errors persist and accumulate, they may develop into more significant breach or lost of trust. Therefore, longitudinal and in-the-wild HRI studies are required to better understand how a robot's errors influence a human-robot team.

In Olivaw's scenario, as a receptionist with access to private information, its violation of others' privacy (E5.2) can lead to significant degradation of users' trust towards it [169]. Degraded users' trust may also reduce the perceived credibility of information provided by Olivaw, as suggested by Burgoon et al. [20] and Kidd and Breazeal [85].

In Calvin's scenario, Frank's trust towards it is the foundation of companionship [102]. Moreover, Calvin's ability to inspire trust in Frank's carers or therapists (E5.6 failures in inspiring trust) is important for collaborating with them and providing an effective, multi-facet support for Frank. During Frank's interaction with other people, such as at the book-reading club, erratic behaviors of Calvin damaging to Frank's social reputation (E5.3) may lead to him feeling embarrassed and becoming less willing to interact with others. Such reduced social interaction can induce feelings of isolation and accelerate Frank's cognitive decline [50].

As a personal robot for medical support, Calvin needs to provide humane and ethical health care and avoid potential harm to the physical or mental well-being of its user [161]. Violation of privacy is a major ethical concern in robot-assisted health care and companion robots, because it may result in users feeling a loss of control and personal liberty [169]. However, it may be necessary for Calvin to release private information under special circumstances, such as providing recordings to Frank's therapist for evaluating the effectiveness of his treatment. The decision of whether to release such information or not should be made in a case-by-case manner. To increase Frank's feeling of empowerment and being in control, consent from him on releasing such information is desirable when possible.

In Reventlov's scenario, when achieving a goal requires collaborative efforts, its ability to enlist collaborators (E5.1) will have direct impact on the task outcome. In such a human-robot team, breaches to reciprocity and fairness can result in conflicts within the team [79]. For high-risk tasks at the polar station, such as search-and-rescue, failures to inspire and reciprocate trust can be especially damaging [156]. In collaboration where knowledge and resources are shared, one's social reputation can have significant impact on how others perceive the quality of knowledge and resources provided by this person [95]. Thus, if Reventlov fails to establish a social reputation of itself during long-term collaboration (E5.4), its teammates may rely less on its support. In current research on multi-agent systems with constrained resources, optimal resource allocation has been the main objective [33]. However, a collaborative team's performance depends largely on people's interpersonal orientation [179], including reciprocity and fairness. Thus, if Reventlov proposes plans that are perceived as unfair, such as allocating 90% usage time of the only computer with GPU to one highest performing individual (E5.5 failures in maintaining fairness), other researchers may not follow its proposals.

4.6 Example use of the proposed taxonomy

In previous sections, we discussed the potential impact of each social error situated in three representative HRI scenarios and existing HRI studies. To illustrate how the proposed taxonomy in Table 4 can help researchers conduct a systematic analysis, here we return to the example at the start of this paper and investigate social errors that occurred during the HRI episode between David and Churi.

Churi is likely to have based its dialogue system on a small set of fixed questions and expected interaction scenarios. David's emotions such as excitement or frustration made no difference to the robot (E1.1 absent emotion recognition, E1.4 unresponsive to user's emotion, and E1.5 absent entrainment response). Its synthetic spoken responses lack expressiveness (E1.2). Churi showed limited social skills (E2), except for adhering to the social routine of reciprocating greetings (E2.6). It misunderstood David's intentions and anticipations, or made no effort to interpret such information (E3). Churi has limited conversational functions. In particular, it synthesized irrelevant responses (E4.2) at misaligned timing (E4.4) with absent non-verbal expressions (E4.3). This broke the dialogue flow and disengaged David (E4.5). No adaption or personalization seems to be included in Churi's design (E4.5). During this specific interaction the collaborative dimension (E5) was not involved, although we can predict that it is unlikely for Churi to inspire effective collaboration in a human-robot team.

At the start, David was amused by some of the errors made by Churi. However, as time progressed, he grew less tolerant of its errors. It is not hard to imagine that soon David will abandon his personal assistant robot entirely, even though he started with high expectation of the interaction experience. Churi the robot has failed to provide satisfying service to its user, let alone establishing an enjoyable social relationship. In fact, in 2019, the Henn-na Hotel laid off half of its 243 robot "staffs", including Churi [52]. To keep the hotel functioning, a large number of tasks had to be redirected to human staff. This potentially undermines the hotel's branding effort of being "the Ultimate in Efficiency".

5 DISCUSSION

In this work, we proposed to classify errors in HRI as performance errors and social errors to separate the functional and social-emotional aspects of HRI. We reviewed human social interactions and existing HRI error studies. This exposed that the social-emotional aspects of HRI, such as emotional interaction, have been overlooked in current research, even though they were shown to have significant influences in human-human interactions. There has been no existing HRI error taxonomy with a user-centered perspective addressing the social-emotional dimension of HRI. Thus, we focused on social errors in HRI and proposed a taxonomy inspired by major attributes of human perception of socio-affective competence and social relationship.

We discussed possible impacts of each type of social error on user experiences and HRI outcomes situated in three HRI scenarios, namely a receptionist robot, a dementia care robot, and a polar expedition support robot, as well as in existing HRI studies. Our discussion highlighted the critical impact perceived socio-affective competence has on HRI: In short-term HRI, social errors of a robot can lead to disengagement, task failures, unsatisfactory services, or compromised collaboration efficiency. For instance, in a human-multi-robot collaboration study [185], the human-robot team had significantly lower task success rate and took longer to achieve the collaborative goal when the robots exchanged information in a covert way, compared to when emotional expressions were displayed during robot-robot communication. In longitudinal HRI, social errors can damage a user's perceived social relationship with a robot and lead to its abandonment. For example, in a study on a home assistant robot [32], by the end of six months, only 47 out of 168 participants indicated that they were still using the robot and intended to continue using it, i.e., an abandonment rate of 72%, even though they were allowed to keep the robot as a compensation for their participation in the study.

Our distinction of performance errors and social errors in HRI aligns with previous social science studies which identify competence and warmth as the two fundamental and universal dimensions which humans base their judgments of individuals and groups upon, i.e., the Stereotype Content Model (SCM) [48]. People's impression of the competence and warmth of others has been shown to elicit specific emotional responses and distinct behavioral tendencies known as the Behaviors from Intergroup Affect and Stereotypes (BIAS) Map [28]. For example, high competence and low warmth can elicit envy and cause harming actions, while low competence and high warmth can elicit pity and cause helping actions. Such mappings between impressions, emotional responses, and behavioral tendencies have been observed in human's interaction with virtual agents [121] and robots [114]. Moreover, first impressions established during short-term HRI were found to persist in repeated interaction sessions [130].

Following SCM and BIAS Map, we expect performance errors to degrade a user's impressions of a robot's competence, while social errors will degrade a user's impressions of a robot's warmth, and result in similar emotional and behavioral reactions to the robot as they have towards humans. Therefore, improving the perceived socio-affective competence of a robot to establish a high warmth impression is key to elicit facilitation and forgiveness in its users. In a number of social robotic studies, humans were found to exhibit bullying or destructive behaviors towards robots [12]. Increasing perceived socio-affective competence of a robot may also encourage empathic perceptions towards it in users, which can potentially reduce destructive behaviors and improve the social acceptance of the robot. For example, Connolly et al. [27] found that people were more likely to intervene in robot mistreatment when it displayed emotional expressions.

5.1 Future directions

Our work has focused on defining social errors in HRI and understanding their potential impact. Because of the underlying differences between human-human interaction and human-robot interaction [99], we encourage researchers to conduct HRI experiments and validate the proposed taxonomy, especially in longitudinal, in-the-wild conditions. For example, to gather initial assessments on human perception of a particular social error, researchers may start with simulated studies. This allows iterative refinement of the HRI scenario design and exposes potential ethical concerns. Because self-reported assessments may contain summative or recall biases [167], it will be important for researchers to implement the refined social error scenarios in HRI experiments and conduct observational and objective assessments in addition to subjective evaluations, for example, using a Pepper robot [134]. Beyond understanding social errors' impacts, in longitudinal HRI studies, it will also be interesting to investigate effective error handling strategies in relation to personalization and adaption of HRI systems.

5.2 Contributions

The proposed taxonomy of social errors in HRI supplies a structured approach with a multi-disciplinary basis to analyze the perceived socio-affective competence of robots. It encourages HRI researchers to engage in a user-centered design of future robots and HRI systems, and to analyze the influence of social errors that occur during interaction. Understanding social errors in HRI is key to design advanced error handling and failure recovery strategies, which leads to better interaction experiences and outcomes. Social errors are also opportunities for achieving personalization and adaption in longitudinal HRI. For example, when a robot is having low confidence in recognizing its user's emotional states, it may explain the situation and the cause of low recognition confidence to the user, and request clarification to improve the accuracy of its emotion recognition model for this specific user. Such error handling with a goal of improving personalization aligns with a recently proposed design guideline of human-AI interaction validated by large scale user studies [3].

Beyond proposing a taxonomy for analyzing social errors and perceived socio-affective competence of HRI systems, our research sheds light on ethical concerns of HRI designs. In particular, how an HRI system may influence the emotions and well-being of its users. To better serve its purposes and avoid potential harm to users, HRI systems should be designed to be diverse and inclusive, whether for various mental health conditions, physical abilities, cultural backgrounds, or other user traits. It is also critical for researchers to understand how an HRI system may influence its users' behaviors, decision-makings, or personal and social lives, and avoid misuse of such influences. Thus, beyond designing HRI systems to be more intelligent and socially acceptable, it is equally important to empower users to take control and make informed and well-supported decisions.

ACKNOWLEDGMENTS

We would like to thank Chathurika Palliya Guruge, Dana Kulić, Dewi Tojib, Frits de Nijs, Jarrod Knibbe, the anonymous reviewers and the editorial board for their valuable feedback that greatly improved this manuscript.

REFERENCES

[1] Carole Adam and Benoit Gaudou. 2016. BDI agents in social simulations: a survey. *The Knowledge Engineering Review* 31, 3 (2016), 207–238.

[2] Jason Aimone, Sheryl Ball, and Brooks King-Casas. 2015. The betrayal aversion elicitation task: An individual level betrayal aversion measure. *PloS one* 10, 9 (2015), e0137491.

[3] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 ACM CHI Conference on Human Factors in Computing Systems*. ACM.

[4] James Andreoni and B Douglas Bernheim. 2009. Social image and the 50-50 norm: A theoretical and experimental analysis of audience effects. *Econometrica* 77, 5 (2009), 1607–1636.

[5] Mohd Anwar and Jim Greer. 2012. Facilitating trust in privacy-preserving e-learning environments. *IEEE Transactions on Learning Technologies* 5, 1 (2012), 62–73.

[6] Michael Argyle and Janet Dean. 1965. Eye-contact, distance and affiliation. *Sociometry* (1965), 289–304.

[7] R. Asaad Baksh, Sharon Abrahams, Bonnie Auyeung, and Sarah E. MacPherson. 2018. The Edinburgh Social Cognition Test (ESCoT): Examining the effects of age on a new measure of theory of mind and social norm understanding. *PLoS ONE* 13, 4 (2018), 1–16.

[8] Simon Baron-Cohen. 1996. Reading the mind in the face: A cross-cultural and developmental study. *Visual Cognition* 3, 1 (1996), 39–60.

[9] Simon Baron-Cohen, Michelle O'Riordan, Rosie Jones, Valerie Stone, and Kate Plaisted. 1999. A new test of social sensitivity: Detection of faux pas in normal children and children with Asperger syndrome. *Journal of Autism and Developmental Disorders* 29, 5 (1999), 407–418.

[10] Simon Baron-cohen and Sally Wheelwright. 2004. The Empathy Quotient: An Investigation of Adults with Asperger Syndrome or High Functioning Autism, and Normal Sex Differences. *Journal of Autism and Developmental Disorders* 34, 2 (2004), 163–175.

[11] Lisa Feldman Barrett. 2017. *How emotions are made: The secret life of the brain.* Houghton Mifflin Harcourt.

[12] Christoph Bartneck, Marcel Verbunt, Omar Mubin, and Abdullah Al Mahmud. 2007. To kill a mockingbird robot. In *Proceedings of the ACM/IEEE international conference on Human-robot interaction*. ACM, 81–87.

[13] Andrea Bauer, Dirk Wollherr, and Martin Buss. 2008. Human-robot collaboration: A survey. *International Journal of Humanoid Robotics* 5, 01 (2008), 47–66.

[14] Matthew I Beane. 2020. In Storage, Yet on Display: An Empirical Investigation of Robots' Value as Social Signals. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 83–91.

[15] Casey C Bennett, Selma Sabanovic, Jennifer A Piatt, Shinichi Nagata, Lori Eldridge, and Natasha Randall. 2017. A robot a day keeps the blues away. In *2017 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE, 536–540.

[16] Joyce Berg, John Dickhaut, and Kevin McCabe. 1995. Trust, reciprocity, and social history. *Games and economic behavior* 10, 1 (1995), 122–142.

[17] David P Biros, Mark Daly, and Gregg Gunsch. 2004. The influence of task load and automation trust on deception detection. *Group Decision and Negotiation* 13, 2 (2004), 173–189.

[18] Cynthia Breazeal, Cory D Kidd, Andrea Lockerd Thomaz, Guy Hoffman, and Matt Berlin. 2005. Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In *2005 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 708–713.

[19] Daniel J Brooks. 2017. *A human-centric approach to autonomous robot failures*. Ph.D. Dissertation. University of Massachusetts Lowell.

[20] Judee K Burgoon, Joseph A Bonito, Bjorn Bengtsson, Carl Cederberg, Magnus Lundeberg, and Lisa Allspach. 2000. Interactivity in human–computer interaction: A study of credibility, understanding, and influence. *Computers in human behavior* 16, 6 (2000), 553–574.

[21] Brian L Burke, Christopher W Dunn, David C Atkins, and Jerry S Phelps. 2004. The emerging evidence base for motivational interviewing: A meta-analytic and qualitative inquiry. *Journal of Cognitive Psychotherapy* 18, 4 (2004), 309–322.

[22] Susan B Campbell, Nina B Leezenbaum, Amanda S Mahoney, Elizabeth L Moore, and Celia A Brownell. 2016. Pretend play and social engagement in toddlers at high and low genetic risk for autism spectrum disorder. *Journal of autism and developmental disorders* 46, 7 (2016), 2305–2316.

[23] Jennifer Carlson and Robin R Murphy. 2005. How UGVs physically fail in the field. *IEEE Transactions on robotics* 21, 3 (2005), 423–437.

[24] David Cesarini, Christopher T Dawes, James H Fowler, Magnus Johannesson, Paul Lichtenstein, and Björn Wallace. 2008. Heritability of cooperative behavior in the trust game. *Proceedings of the National Academy of sciences* 105, 10 (2008), 3721–3726.

[25] Nikhil Churamani, Paul Anton, Marc Brügger, Erik Fließwasser, Thomas Hummel, Julius Mayer, Waleed Mustafa, Hwei Geok Ng, Thi Linh Chi Nguyen, Quan Nguyen, et al. 2017. The Impact of Personalisation on Human-Robot Interaction in Learning Scenarios. In *Proceedings of the 5th International Conference on Human Agent Interaction*. ACM, 171–180.

[26] Sheldon Cohen. 2004. Social relationships and health. *American psychologist* 59, 8 (2004), 676.

[27] Joe Connolly, Viola Mocz, Nicole Salomons, Joseph Valdez, Nathan Tsoi, Brian Scassellati, and Marynel Vázquez. 2020. Prompting Prosocial Human Interventions in Response to Robot Mistreatment. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 211–220.

[28] Amy JC Cuddy, Susan T Fiske, and Peter Glick. 2007. The BIAS map: behaviors from intergroup affect and stereotypes. *Journal of personality and social psychology* 92, 4 (2007), 631.

[29] Benjamin MP Cuff, Sarah J Brown, Laura Taylor, and Douglas J Howat. 2016. Empathy: A review of the concept. *Emotion Review* 8, 2 (2016), 144–153.

[30] Mark H Davis. 1980. *Interpersonal reactivity index*. Edwin Mellen Press.

[31] Mark H Davis et al. 1980. A multidimensional approach to individual differences in empathy. (1980).

[32] Maartje De Graaf, Somaya Ben Allouch, and Jan Van Dijk. 2017. Why do they refuse to use my robot?: Reasons for non-use derived from a long-term home study. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 224–233.

[33] Frits de Nijs. 2019. *Resource-constrained Multi-agent Markov Decision Processes*. Ph.D. Dissertation. Delft University of Technology, SIKS Dissertation Series.

[34] Jean Decety and Philip L Jackson. 2004. The functional architecture of human empathy. *Behavioral and cognitive neuroscience reviews* 3, 2 (2004), 71–100.

[35] Jean Decety and Meghan Meyer. 2008. From emotion resonance to empathic understanding: A social developmental neuroscience account. *Development and psychopathology* 20, 4 (2008), 1053–1080.

[36] Lambert Deckers and John Devine. 1981. Humor by violating an existing expectancy. *The Journal of Psychology* 108, 1 (1981), 107–110.

[37] Graham Denyer. 2005. *Davis Station Antarctica, Australian Antarctic Programme.* Wikipedia. https://en.wikipedia.org/wiki/Davis_Station

[38] Pieter TM Desmet, David De Cremer, and Eric van Dijk. 2011. Trust recovery following voluntary or forced financial compensations in the trust game: The role of trait forgiveness. *Personality and Individual Differences* 51, 3 (2011), 267–273.

[39] Joshua J Diehl, Charles R Crowell, Michael Villano, Kristin Wier, Karen Tang, and Laurel D Riek. 2014. Clinical applications of robots in autism spectrum disorder diagnosis and treatment. In *Comprehensive guide to autism.* Springer, 411–422.

[40] Andreas Diekmann. 2004. The power of reciprocity: Fairness, reciprocity, and stakes in variants of the dictator game. *Journal of conflict resolution* 48, 4 (2004), 487–505.

[41] Pierre Dillenbourg and David Traum. 2006. Sharing solutions: Persistence and grounding in multimodal collaborative problem solving. *The Journal of the Learning Sciences* 15, 1 (2006), 121–151.

[42] Grazia D'Onofrio, Daniele Sancarlo, Francesco Ricciardi, Francesco Panza, Davide Seripa, Filippo Cavallo, Francesco Giuliani, and Antonio Greco. 2017. Information and communication technologies for the activities of daily living in older patients with dementia: A systematic review. *Journal of Alzheimer's Disease* 57, 3 (2017), 927–935.

[43] Isabel Dziobek, Stefan Fleck, Elke Kalbe, Kimberley Rogers, Jason Hassenstab, Matthias Brand, Josef Kessler, Jan K Woike, Oliver T Wolf, and Antonio Convit. 2006. Introducing MASC: A movie for the assessment of social cognition. *Journal of autism and developmental disorders* 36, 5 (2006), 623–636.

[44] Jon Elster. 1989. Social norms and economic theory. *Journal of economic perspectives* 3, 4 (1989), 99–117.

[45] Ernst Fehr and Urs Fischbacher. 2004. Social norms and human cooperation. *Trends in cognitive sciences* 8, 4 (2004), 185–190.

[46] Oriel FeldmanHall and Luke J Chang. 2018. Social Learning: Emotions Aid in Optimizing Goal-Directed Social Behavior. In *Goal-Directed Decision Making.* Elsevier, 309–330.

[47] Gary Alan Fine and Ugo Corte. 2017. Group pleasures: Collaborative commitments, shared narrative, and the sociology of fun. *Sociological Theory* 35, 1 (2017), 64–86.

[48] Susan T Fiske, Amy JC Cuddy, Peter Glick, and Jun Xu. 2002. A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. *Journal of personality and social psychology* 82, 6 (2002), 878.

[49] Terrence Fong, Charles Thorpe, and Charles Baur. 2003. Collaboration, dialogue, human-robot interaction. In *Robotics Research.* Springer, 255–266.

[50] Laura Fratiglioni, Hui-Xin Wang, Kjerstin Ericsson, Margaret Maytan, and Bengt Winblad. 2000. Influence of social network on occurrence of dementia: A community-based longitudinal study. *The lancet* 355, 9212 (2000), 1315–1319.

[51] Uta Frith. 2003. *Autism: Explaining the enigma.* Blackwell Publishing.

[52] Alastair Gale and Takashi Mochizuki. 2019. Robot Hotel Loses Love for Robots. *The Wall Street Journal* (Jan 2019). https://www.wsj.com/articles/robot-hotel-loses-love-for-robots-11547484628

[53] Jolene Galegher, Robert E Kraut, and Carmen Egido. 2014. *Intellectual teamwork: Social and technological foundations of cooperative work.* Psychology Press.

[54] Denise Y. Geiskkovitch, Raquel Thiessen, James E. Young, and Melanie R. Glenwright. 2019. What? That's Not a Chair!: How Robot Informational Errors Affect Children's Trust Towards Robots. In *Proceedings of the 2019 ACM/IEEE International Conference on Human-Robot Interaction.* ACM, 48–56.

[55] Andrew D Gershoff and Jonathan J Koehler. 2011. Safety first? The role of emotion in safety product betrayal aversion. *Journal of Consumer Research* 38, 1 (2011), 140–150.

[56] Manuel Giuliani, Nicole Mirnig, Gerald Stollnberger, Susanne Stadler, Roland Buchner, and Manfred Tscheligi. 2015. Systematic analysis of video data from different human-robot interaction studies: A categorization of social signals during error situations. *Frontiers in psychology* 6 (2015), 931.

[57] Ofer Golan, Simon Baron-Cohen, Jacqueline J Hill, and Yael Golan. 2006. The "reading the mind in films" task: Complex emotion recognition in adults with and without autism spectrum conditions. *Social Neuroscience,* 1, 2 (2006), 111–123.

[58] Susel Góngora Alonso, Sofiane Hamrioui, Isabel de la Torre Díez, Eduardo Motta Cruz, Miguel López-Coronado, and Manuel Franco. 2018. Social Robots for People with Aging and Dementia: A Systematic Review of Literature. *Telemedicine and e-Health* (2018).

[59] Katherine Gotham, Susan Risi, Andrew Pickles, and Catherine Lord. 2007. The Autism Diagnostic Observation Schedule: Revised algorithms for improved diagnostic validity. *Journal of autism and developmental disorders* 37, 4 (2007), 613.

[60] Cornelius T Gross and Newton Sabino Canteras. 2012. The many paths to fear. *Nature Reviews Neuroscience* 13, 9 (2012), 651.

[61] Shang Guo, Jonathan Lenchner, Jonathan Connell, Mishal Dholakia, and Hidemasa Muta. 2017. Conversational bootstrapping and other tricks of a concierge robot. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 73–81.

[62] Amy G Halberstadt, Susanne A Denham, and Julie C Dunsmore. 2001. Affective social competence. *Social development* 10, 1 (2001), 79–119.

[63] Pippa Hall. 2005. Interprofessional teamwork: Professional cultures as barriers. *Journal of Interprofessional care* 19, sup1 (2005), 188–196.

[64] Peter A Hancock, Deborah R Billings, Kristin E Schaefer, Jessie YC Chen, Ewart J De Visser, and Raja Parasuraman. 2011. A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors* 53, 5 (2011), 517–527.

[65] Elaine Hatfield, John T Cacioppo, and Richard L Rapson. 1993. Emotional contagion. *Current directions in psychological science* 2, 3 (1993), 96–100.

[66] Lisa Heavey, Wendy Phillips, Simon Baron-Cohen, and Michael Rutter. 2000. The Awkward Moments Test: A naturalistic measure of social understanding in autism. *Journal of autism and developmental disorders* 30, 3 (2000), 225–236.

[67] Marcel Heerink, Ben Krose, Vanessa Evers, and Bob Wielinga. 2006. The influence of a robot's social abilities on acceptance by elderly users. In *ROMAN 2006-The 15th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 521–526.

[68] Julia Hirschberg. 2011. Speaking more like you: Entrainment in conversational speech. In *Twelfth Annual Conference of the International Speech Communication Association*.

[69] Guy Hoffman and Cynthia Breazeal. 2004. Collaboration in human-robot teams. In *AIAA 1st Intelligent Systems Technical Conference*. 6434.

[70] Shanee Sarah Honig and Tal Oron-Gilad. 2018. Understanding and Resolving Failures in Human-Robot Interaction: Literature Review and Model Development. *Frontiers in psychology* 9 (2018), 861.

[71] Leonard M Horowitz, Saul E Rosenberg, Barbara A Baer, Gilbert Ureño, and Valerie S Villaseñor. 1988. Inventory of interpersonal problems: Psychometric properties and clinical applications. *Journal of consulting and clinical psychology* 56, 6 (1988), 885.

[72] James S House, Cynthia Robbins, and Helen L Metzner. 1982. The association of social relationships and activities with mortality: Prospective evidence from the Tecumseh Community Health Study. *American journal of epidemiology* 116, 1 (1982), 123–140.

[73] Chien-Ming Huang and Bilge Mutlu. 2014. Learning-based modeling of multimodal behaviors for humanlike robots. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*. ACM, 57–64.

[74] Bahar Irfan, Aditi Ramachandran, Samuel Spaulding, Dylan F Glas, Iolanda Leite, and Kheng Lee Koay. 2019. Personalization in Long-Term Human-Robot Interaction. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 685–686.

[75] K. Izuma, K. Matsumoto, C. F. Camerer, and R. Adolphs. 2011. Insensitivity to social reputation in autism. *Proceedings of the National Academy of Sciences* 108, 42 (2011), 17302–17307.

[76] John A Johnson, Jonathan M Cheek, and Robert Smither. 1983. The structure of empathy. *Journal of personality and social psychology* 45, 6 (1983), 1299–1312.

[77] Gareth R Jones and Jennifer M George. 1998. The experience and evolution of trust: Implications for cooperation and teamwork. *Academy of management review* 23, 3 (1998), 531–546.

[78] Malte F Jung. 2017. Affective Grounding in Human-Robot Interaction. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 263–273.

[79] Malte F Jung, Nikolas Martelaro, and Pamela J Hinds. 2015. Using robots to moderate team conflict: the case of repairing violations. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 229–236.

[80] Daniel Kahneman, Jack L Knetsch, and Richard Thaler. 1986. Fairness as a constraint on profit seeking: Entitlements in the market. *The American economic review* (1986), 728–741.

[81] Daniel Kahneman, David Schkade, and Cass Sunstein. 1998. Shared outrage and erratic awards: The psychology of punitive damages. *Journal of Risk and Uncertainty* 16, 1 (1998), 49–86.

[82] Yusuke Kato, Takayuki Kanda, and Hiroshi Ishiguro. 2015. May I help you?: Design of human-like polite approaching behavior. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 35–42.

[83] Richard Kelley, Alireza Tavakkoli, Christopher King, Monica Nicolescu, Mircea Nicolescu, and George Bebis. 2008. Understanding human intentions via hidden Markov models in autonomous mobile robots. In *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*. ACM, 367–374.

[84] Adam Kendon. 1990. *Conducting interaction: Patterns of behavior in focused encounters*. Vol. 7. CUP Archive.

[85] Cory D Kidd and Cynthia Breazeal. 2004. Effect of a robot on user perceptions. In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566)*, Vol. 4. IEEE, 3559–3564.

[86] Charles E Kimble and Steven D Seidel. 1991. Vocal signs of confidence. *Journal of Nonverbal Behavior* 15, 2 (1991), 99–105.

[87] Erik O. Kimbrough and Alexander Vostroknutov. 2016. Norms Make Preferences Social. *Journal of the European Economic Association* 14, 3 (2016), 608–638.

[88] Masanori Kimura and Ikuo Daibo. 2006. Interactional synchrony in conversations about emotional episodes: A measurement by "the between-participants pseudosynchrony experimental paradigm". *Journal of Nonverbal Behavior* 30, 3 (2006), 115–126.

[89] Kyveli Kompatsiari, Vadim Tikhanoff, Francesca Ciardo, Giorgio Metta, and Agnieszka Wykowska. 2017. The importance of mutual gaze in human-robot interaction. In *International Conference on Social Robotics*. Springer, 443–452.

[90] Dimosthenis Kontogiorgos, Andre Pereira, Boran Sahindal, Sanne van Waveren, and Joakim Gustafson. 2020. Behavioural Responses to Robot Conversational Failures. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 53–62.

[91] Agnieszka Korchut, Sebastian Szklener, Carla Abdelnour, Natalia Tantinya, Joan Hernández-Farigola, Joan Carles Ribes, Urszula Skrobas, Katarzyna Grabowska-Aleksandrowicz, Dorota Szczęśniak-Stańczyk, and Konrad Rejdak. 2017. Challenges for service robots—requirements of elderly adults with cognitive impairments. *Frontiers in neurology* 8 (2017), 228.

[92] Alyssa Kubota, Emma IC Peterson, Vaishali Rajendren, Hadas Kress-Gazit, and Laurel D Riek. 2020. JESSIE: Synthesizing Social Robot Behaviors for Personalized Neurorehabilitation and Beyond. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 121–130.

[93] Sari Kujala. 2003. User involvement: a review of the benefits and challenges. *Behaviour & information technology* 22, 1 (2003), 1–16.

[94] JC Laprie. 1995. DEPENDABLE COMPUTING AND FAULT TOLERANCE: CONCEPTS AND TERMINOLOGY. In *Twenty-Fifth International Symposium on Fault-Tolerant Computing, Highlights from Twenty-Five Years*. IEEE, 2.

[95] Nada Lavrac, Peter Ljubic, Tanja Urbancic, Gregor Papa, Mitja Jermol, and Stefan Bollhalter. 2007. Trust modeling for networked organizations using reputation and collaboration estimates. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 37, 3 (2007), 429–439.

[96] John O Ledyard. 1994. Public goods: A survey of experimental research. (1994).

[97] Min Kyung Lee, Sara Kiesler, Jodi Forlizzi, Siddhartha Srinivasa, and Paul Rybski. 2010. Gracefully mitigating breakdowns in robotic services. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 203–210.

[98] Wendy Leeds-Hurwitz. 1987. The social history of the natural history of an interview: A multidisciplinary investigation of social communication. *Research on Language and Social Interaction* 20, 1-4 (1987), 1–51.

[99] Benedikt Leichtmann and Verena Nitsch. 2020. How much distance do humans keep toward robots? Literature review, meta-analysis, and theoretical considerations on personal space in human-robot interaction. *Journal of Environmental Psychology* 68 (2020), 101386.

[100] Sergey Levine, Peter Pastor, Alex Krizhevsky, Julian Ibarz, and Deirdre Quillen. 2018. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International Journal of Robotics Research* 37, 4-5 (2018), 421–436.

[101] Rivka Levitan, Agustín Gravano, Laura Willson, Stefan Benus, Julia Hirschberg, and Ani Nenkova. 2012. Acoustic-prosodic entrainment and social behavior. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human language*

technologies. Association for Computational Linguistics, 11–19.

[102] J David Lewis and Andrew Weigert. 1985. Trust as a social reality. *Social forces* 63, 4 (1985), 967–985.

[103] M. Lewis, JM Laviland-Jones, and L Fildman Barret. 2012. Self-Conscious Emotions-Embarrassment, Pride, Shame, and Guilt. *Handbook of Emotions* (2012).

[104] Chang Liu, Jessica B Hamrick, Jaime F Fisac, Anca D Dragan, J Karl Hedrick, S Shankar Sastry, and Thomas L Griffiths. 2016. Goal inference improves objective and perceived performance in human-robot collaboration. In *Proceedings of the 2016 international conference on autonomous agents & multiagent systems*. International Foundation for Autonomous Agents and Multiagent Systems, 940–948.

[105] Catherine Lord, Susan Risi, Linda Lambrecht, Edwin H Cook, Bennett L Leventhal, Pamela C DiLavore, Andrew Pickles, and Michael Rutter. 2000. The Autism Diagnostic Observation Schedule—Generic: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of autism and developmental disorders* 30, 3 (2000), 205–223.

[106] Maurice Lorr, Richard P Youniss, and Edward C Stefic. 1991. An inventory of social skills. *Journal of personality assessment* 57, 3 (1991), 506–520.

[107] Joseph Lyons, Charlene Stokes, David Garcia, Justin Adams, and Dave Ames. 2009. Trust and decision-making: An empirical platform (CCP 204). *IEEE Aerospace and Electronic Systems Magazine* 24, 10 (2009), 36–41.

[108] Michael A Madaio, Rae Lasko, Justine Cassell, and Amy Ogan. 2017. Using Temporal Association Rule Mining to Predict Dyadic Rapport in Peer Tutoring. In *Proceedings of the 10th International Conference on Educational Data Mining*.

[109] Bertram F Malle and Stuti Thapa Magar. 2017. What kind of mind do I want in my robot?: Developing a measure of desired mental capacities in social robots. In *Proceedings of the companion of the 2017 ACM/IEEE international conference on human-robot interaction*. ACM, 195–196.

[110] Stacy C Marsella and Jonathan Gratch. 2009. EMA: A process model of appraisal dynamics. *Cognitive Systems Research* 10, 1 (2009), 70–90.

[111] Andrew Maul. 2012. The validity of the Mayer-Salovey-Caruso Emotional Intelligence Test (MSCEIT) as a measure of emotional intelligence. *Emotion Review* 4, 4 (2012), 394–402.

[112] Maranda McBride and Shona Morgan. 2010. Trust calibration for automated decision aids. *Institute for Homeland Security Solutions* (2010).

[113] Albert Mehrabian and Norman Epstein. 1972. A measure of emotional empathy. *Journal of personality* 40, 4 (1972), 525–543.

[114] Hannah Mieczkowski, Sunny Xun Liu, Jeffrey Hancock, and Byron Reeves. 2019. Helping Not Hurting: Applying the Stereotype Content Model and BIAS Map to Social Robotics. In *Proceedings of the 2019 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 222–229.

[115] Nicole Mirnig, Manuel Giuliani, Gerald Stollnberger, Susanne Stadler, Roland Buchner, and Manfred Tscheligi. 2015. Impact of robot actions on social signals and reaction times in HRI error situations. In *International Conference on Social Robotics*. Springer, 461–471.

[116] Nicole Mirnig, Gerald Stollnberger, Markus Miksch, Susanne Stadler, Manuel Giuliani, and Manfred Tscheligi. 2017. To err is robot: How humans assess and act toward an erroneous social robot. *Frontiers in Robotics and AI* 4 (2017), 21.

[117] Thomas M Moerland, Joost Broekens, and Catholijn M Jonker. 2018. Emotion in reinforcement learning agents and robots: A survey. *Machine Learning* 107, 2 (2018), 443–480.

[118] Lorenza Mondada. 2016. Challenges of multimodality: Language and the body in social interaction. *Journal of Sociolinguistics* 20, 3 (2016), 336–366.

[119] Elaine Mordoch, Angela Osterreicher, Lorna Guse, Kerstin Roger, and Genevieve Thompson. 2013. Use of social commitment robots in the care of elderly people with dementia: A literature review. *Maturitas* 74, 1 (2013), 14–20.

[120] Shrikanth Narayanan and Panayiotis G Georgiou. 2013. Behavioral signal processing: Deriving human behavioral informatics from speech and language. *Proc. IEEE* 101, 5 (2013), 1203–1233.

[121] Truong-Huy D Nguyen, Elin Carstensdottir, Nhi Ngo, Magy Seif El-Nasr, Matt Gray, Derek Isaacowitz, and David Desteno. 2015. Modeling warmth and competence in virtual characters. In *International Conference on Intelligent Virtual Agents*. Springer, 167–180.

[122] Paula M Niedenthal, Magdalena Rychlowska, Adrienne Wood, and Fangyun Zhao. 2018. Heterogeneity of long-history migration predicts smiling, laughter and positive emotion across the globe and within the United States. *PloS one* 13, 8 (2018), e0197651.

[123] Stefanos Nikolaidis, Yu Xiang Zhu, David Hsu, and Siddhartha Srinivasa. 2017. Human-robot mutual adaptation in shared autonomy. In *Proceedings of the 2017 ACM/IEEE International Conference on*

Human-Robot Interaction. ACM, 294–302.

[124] Donald A Norman and Stephen W Draper. 1986. *User centered system design: New perspectives on human-computer interaction*. CRC Press.

[125] Carlo Nuccio, Agnese Augello, Salvatore Gaglio, and Giovanni Pilato. 2017. Interaction Capabilities of a Robotic Receptionist. In *International Conference on Intelligent Interactive Multimedia Systems and Services*. Springer, 171–180.

[126] Keith Oatley and Philip N Johnson-Laird. 1987. Towards a cognitive theory of emotions. *Cognition and emotion* 1, 1 (1987), 29–50.

[127] Anna Ogarkova. 2016. Translatability of emotions. In *Emotion measurement*. Elsevier, 575–599.

[128] Sharon Oviatt. 2013. *The design of future educational interfaces*. Routledge.

[129] Sharon Oviatt, Margaret MacEachern, and Gina-Anne Levow. 1998. Predicting hyperarticulate speech during human-computer error resolution. *Speech Communication* 24, 2 (1998), 87–110.

[130] Maike Paetzel, Giulia Perugia, and Ginevra Castellano. 2020. The Persistence of First Impressions: The Effect of Repeated Interactions on the Perception of a Social Robot. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 73–82.

[131] Ana Paiva, Iolanda Leite, Hana Boukricha, and Ipke Wachsmuth. 2017. Empathy in virtual agents and robots: A survey. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 7, 3 (2017), 11.

[132] Lucas Paletta, Maria Fellner, Sandra Schüssler, Julia Zuschnegg, Josef Steiner, Alexander Lerch, Lara Lammer, and Dimitrios Prodromou. 2018. AMIGO: Towards Social Robot based Motivation for Playful Multimodal Intervention in Dementia. In *Proceedings of the 11th PErvasive Technologies Related to Assistive Environments Conference*. ACM, 421–427.

[133] Lawrence A Palinkas and Peter Suedfeld. 2008. Psychological effects of polar expeditions. *The Lancet* 371, 9607 (2008), 153–163.

[134] Amit Pandey and Rodolphe Gelin. 2018. A Mass-Produced Sociable Humanoid Robot: Pepper, The First Machine of Its Kind. *IEEE Robotics & Automation Magazine* 99 (2018).

[135] Luiz Pessoa. 2018. Emotion and the Interactive Brain: Insights From Comparative Neuroanatomy and Complex Systems. *Emotion Review* 10, 3 (2018), 204–216.

[136] Rosalind W Picard. 2000. *Affective computing*. MIT press.

[137] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion* 37 (2017), 98–125.

[138] KP Rankin. 2008. Social Norms Questionaire. *NINDS: Domain Specific Tasks of Executive Function* (2008).

[139] KP Rankin, E Baldwin, C Pace-Savitsky, JH Kramer, and BL Miller. 2005. Self awareness and personality change in dementia. *Journal of Neurology, Neurosurgery & Psychiatry* 76, 5 (2005), 632–639.

[140] PL Patrick Rau, Ye Li, and Dingjun Li. 2009. Effects of communication style and culture on ability to accept recommendations from robots. *Computers in Human Behavior* 25, 2 (2009), 587–595.

[141] HC Ravichandar and A Dani. 2017. Intention inference for human-robot collaboration in assistive robotics. In *Human Modelling for Bio-Inspired Robotics*. Elsevier, 217–249.

[142] Harish Chaandar Ravichandar and Ashwin P Dani. 2017. Human intention inference using expectation-maximization algorithm with online model learning. *IEEE Transactions on Automation Science and Engineering* 14, 2 (2017), 855–868.

[143] Laurel D Riek. 2016. Robotics technology in mental health care. In *Artificial intelligence in behavioral and mental health care*. Elsevier, 185–203.

[144] Ronald E Riggio. 1986. Assessment of basic social skills. *Journal of Personality and social Psychology* 51, 3 (1986), 649.

[145] Paul Robinette, Ayanna M Howard, and Alan R Wagner. 2017. Effect of Robot Performance on Human-Robot Trust in Time-Critical Situations. *IEEE Transactions on Human-Machine Systems* 47, 4 (2017), 425–436.

[146] Howard J Rosen, Katherine Pace-Savitsky, Richard J Perry, Joel H Kramer, Bruce L Miller, and Robert W Levenson. 2004. Recognition of emotion in the frontal and temporal variants of frontotemporal dementia. *Dementia and geriatric cognitive disorders* 17, 4 (2004), 277–281.

[147] Robert Ross, Rem Collier, and Gregory MP O'Hare. 2004. Demonstrating social error recovery with AgentFactory. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems-Volume 3*. IEEE Computer Society, 1424–1425.

[148] Alessandra Rossi, Kerstin Dautenhahn, Kheng Lee Koay, and Michael L Walters. 2017. Human perceptions of the severity of domestic robot errors. In *International Conference on Social Robotics*.

Springer, 647–656.

[149] Silvia Rossi, Francois Ferland, and Adriana Tapus. 2017. User profiling and behavioral adaptation for HRI: A survey. *Pattern Recognition Letters* 99 (2017), 3–12.

[150] Mary Beth Rosson and John M Carroll. 2009. Scenario based design. *Human-computer interaction. boca raton, FL* (2009), 145–162.

[151] James A Russell. 1991. Culture and the categorization of emotions. *Psychological bulletin* 110, 3 (1991), 426.

[152] Mel D Rutherford, Simon Baron-Cohen, and Sally Wheelwright. 2002. Reading the mind in the voice: A study with normal adults and adults with Asperger syndrome and high functioning autism. *Journal of autism and developmental disorders* 32, 3 (2002), 189–194.

[153] Magdalena Rychlowska, Rachael E Jack, Oliver GB Garrod, Philippe G Schyns, Jared D Martin, and Paula M Niedenthal. 2017. Functional smiles: Tools for love, sympathy, and war. *Psychological science* 28, 9 (2017), 1259–1270.

[154] Selma Sabanovic, Casey C Bennett, Wan-Ling Chang, and Lesa Huber. 2013. PARO robot affects diverse interaction modalities in group sensory therapy for older adults with dementia. In *2013 IEEE International Conference on Rehabilitation Robotics (ICORR)*. IEEE, 1–6.

[155] Alessandra Maria Sabelli, Takayuki Kanda, and Norihiro Hagita. 2011. A conversational robot in an elderly care center: An ethnographic study. In *Human-Robot Interaction (HRI), 2011 6th ACM/IEEE International Conference on*. IEEE, 37–44.

[156] Maha Salem, Gabriella Lakatos, Farshid Amirabdollahian, and Kerstin Dautenhahn. 2015. Would you trust a (faulty) robot?: Effects of error, task type and personality on human-robot cooperation and trust. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 141–148.

[157] Kristin Schaefer. 2013. *The perception and measurement of human-robot trust*. Ph.D. Dissertation. University of Central Florida.

[158] Susanne Scheibe and Laura L Carstensen. 2010. Emotional aging: Recent findings and future trends. *The Journals of Gerontology: Series B* 65, 2 (2010), 135–144.

[159] Sarah Strohkorb Sebo, Priyanka Krishnamurthi, and Brian Scassellati. 2019. "I Don't Believe You": Investigating the Effects of Robot Trust Violation and Repair. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 57–65.

[160] Stela H Seo, Keelin Griffin, James E Young, Andrea Bunt, Susan Prentice, and Verónica Loureiro-Rodríguez. 2018. Investigating people's rapport building and hindering behaviors when working with a collaborative robot. *International Journal of Social Robotics* 10, 1 (2018), 147–161.

[161] Amanda Sharkey and Noel Sharkey. 2012. Granny and the robots: Ethical issues in robot care for the elderly. *Ethics and information technology* 14, 1 (2012), 27–40.

[162] Thomas B Sheridan. 2016. Human–robot interaction: Status and challenges. *Human factors* 58, 4 (2016), 525–532.

[163] Herbert A Simon. 1967. Motivational and emotional controls of cognition. *Psychological review* 74, 1 (1967), 29.

[164] Dirk Sliwka. 2007. Trust as a signal of a social norm and the hidden costs of incentive schemes. *American Economic Review* 97, 3 (2007), 999–1012.

[165] Paul Slovic, Ellen Peters, Melissa L Finucane, and Donald G MacGregor. 2005. Affect, risk, and decision making. *Health psychology* 24, 4S (2005), S35.

[166] Rajka Smiljanic and Rachael C Gilbert. 2017. Acoustics of clear and noise-adapted speech in children, young, and older adults. *Journal of Speech, Language, and Hearing Research* 60, 11 (2017), 3081–3096.

[167] Douglas K Snyder, Richard E Heyman, and Stephen N Haynes. 2005. Evidence-based approaches to assessing couple distress. *Psychological assessment* 17, 3 (2005), 288.

[168] Gerald Steinbauer. 2013. A survey about faults of robots used in RoboCup. In *RoboCup 2012: robot soccer world cup XVI*. Springer, 344–355.

[169] Dag Sverre Syrdal, Michael L Walters, Nuno Otero, Kheng Lee Koay, and Kerstin Dautenhahn. 2007. He knows when you are sleeping-privacy and the personal robot companion. In *Proc. Workshop Human Implications of Human-Robot Interaction, Association for the Advancement of Artificial Intelligence (AAAI'07)*. 28–33.

[170] Peggy A Thoits. 2004. Emotion norms, emotion work, and social order. In *Feelings and emotions: The Amsterdam symposium*. Cambridge University Press New York, NY, 359–378.

[171] Suzanne Tolmeijer, Astrid Weiss, Marc Hanheide, Felix Lindner, Thomas M Powers, Clare Dixon, and Myrthe L Tielman. 2020. Taxonomy of Trust-Relevant Failures and Mitigation Strategies. In

Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction. 3–12.

[172] Ha Trinh, Reza Asadi, Darren Edge, and T Bickmore. 2017. Robocop: A robotic coach for oral presentations. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 2 (2017), 1–24.

[173] Tom Richard Tyler. 1996. *Trust in organizations: Frontiers of theory and research*. Sage.

[174] Anouk van Maris, Hagen Lehmann, Lorenzo Natale, and Beata Grzyb. 2017. The Influence of a Robot's Embodiment on Trust: A Longitudinal Study. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 313–314.

[175] Fabio Vannucci, Alessandra Sciutti, Marco Jacono, Giulio Sandini, and Francesco Rea. 2017. Adaptation to a humanoid robot in a collaborative joint task. In *Robot and Human Interactive Communication (RO-MAN), 2017 26th IEEE International Symposium on*. IEEE, 360–365.

[176] Alessandro Vinciarelli, Maja Pantic, Dirk Heylen, Catherine Pelachaud, Isabella Poggi, Francesca D'Errico, and Marc Schroeder. 2012. Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *IEEE Transactions on Affective Computing* 3, 1 (2012), 69–87.

[177] Ning Wang, David V Pynadath, and Susan G Hill. 2016. Trust calibration within a human-robot team: Comparing automatically generated explanations. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*. IEEE Press, 109–116.

[178] Kristine N Williams, Ruth Herman, Byron Gajewski, and Kristel Wilson. 2009. Elderspeak communication: Impact on dementia care. *American Journal of Alzheimer's Disease & Other Dementias®* 24, 1 (2009), 11–20.

[179] Anita Williams Woolley, Christopher F Chabris, Alex Pentland, Nada Hashmi, and Thomas W Malone. 2010. Evidence for a collective intelligence factor in the performance of human groups. *science* 330, 6004 (2010), 686–688.

[180] Jane Wu, Erin Paeng, Kari Linder, Piercarlo Valdesolo, and James C Boerkoel. 2016. Trust and cooperation in human-robot decision making. In *2016 aaai fall symposium series*.

[181] Bo Xiao, Zac E Imel, Panayiotis Georgiou, David C Atkins, and Shrikanth S Narayanan. 2016. Computational analysis and simulation of empathic behaviors: A survey of empathy modeling with behavioral signal processing framework. *Current psychiatry reports* 18, 5 (2016), 49.

[182] Akihito Yatsuda, Toshiyuki Haramaki, and Hiroaki Nishino. 2018. A Robot Gesture Framework for Watching and Alerting the Elderly. In *International Conference on Network-Based Information Systems*. Springer, 132–143.

[183] Zhou Yu, Vikram Ramanarayanan, Patrick Lange, and David Suendermann-Oeft. 2019. An open-source dialog system with real-time engagement tracking for job interview training applications. In *Advanced Social Interaction with Agents*. Springer, 199–207.

[184] Beste F Yuksel, Penny Collisson, and Mary Czerwinski. 2017. Brains or beauty: How to engender trust in user-agent interactions. *ACM Transactions on Internet Technology (TOIT)* 17, 1 (2017), 2.

[185] Shujie Zhou and Tian Leimin. 2020. Would you help a sad robot? Influence of robots' emotional expressions on human-multi-robot collaboration. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE.

Received 6th June 2019; revised 27th July 2020; accepted 2020